PROC SURVDIFF

by

Thomas R. Fleming, Glenn A. Augustine,
Sharon A. Elcombe, Kenneth P. Offord

Technical Report Series, No. 26

November 9, 1984

PROC SURVDIFF[1]

Thomas R. Fleming[2], Glenn A. Augustine, Sharon A. Elcombe, Kenneth P. Offord[3]

Mayo Clinic

SURVDIFF provides non-parametric statistics to compare survival curves in independent samples. Linear rank test statistics include the Gehan-Wilcoxon (GW)[(1)] and the $G^{\rho,\gamma}$ class of statistics where $\rho \geq 0$, $\gamma \geq 0$. Special cases are the Harrington-Fleming G^{ρ} statistics[(2)] ($\rho \geq 0$, $\gamma = 0$), the log-rank test[(3)] (LR) ($\rho = 0$, $\gamma = 0$) and Peto-Peto-Wilcoxon[(4)] (PPW) ($\rho = 1$, $\gamma = 0$). These linear rank tests have been developed for $r \geq 2$ sample situations, with corresponding one-sample goodness-of-fit statistics defined, as well. For one-sample tests, the $G^{\rho,\gamma}$ is available for $\rho \geq 0$, $\gamma = 0,1,2$. For r=2 sample situations, versions of the $G^{\rho,\gamma}$ and GW test statistics are also available for testing departures from a proportional hazards model where the prespecified proportionality constant need not be unity.

In testing for the equality of two survival curves, a generalized Smirnov[(5)] (GS) test statistic and a class of κ^{ρ} [(6,7)] statistics, $\rho \geq 0$, which are non-linear supremum-type rank statistics, have been included.

Users should view SURVDIFF as a replacement for the SURVTEST procedure. It is the natural complement to the SURVFIT procedure which estimates survival curves.

[1] Development was supported in part by Research Grant CA-24089 from the National Cancer Institute.

[2] Current address: Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195.

[3] Address correspondence to: Kenneth P. Offord, Medical Research Statistics, Mayo Clinic, Rochester, MN 55905.

A. Statistical Development

A1. Notation. Assume the following data are available on the k^{th} individual, k=1, ..., N, where N indicates the total number of individuals over all samples. Let the k^{th} individual's observation time be denoted by X^k and let Δ^k be an event indicator denoting whether the observation time is an event time ($\Delta^k=1$) or a censorship time ($\Delta^k=0$). Also let Z^k be the sample indicator, so $Z^k \epsilon \{0,1,\ldots,r-1\}$. Letting i index the sample, then the number of individuals in the i^{th} sample is given by

$$N_i \equiv \sum_{k=1}^{N} I\{Z^k=i\},$$

where $I\{A\}$ refers to the indicator function for the event A. Note that

$$N = \sum_{i=0}^{r-1} N_i.$$

If we set $Y^k(x)=I\{X^k \geq x\}$, then $Y_i(x) = \sum_{k=1}^{N} Y^k(x)\, I\{Z^k=i\}$ represents the size of the risk set at time x in sample i; i.e., the number of individuals in sample i who are observed for at least x days. Define $N^k(x)=I\{X^k \leq x,\ \Delta^k=1\}$, so $N_i(x)= \sum_{k=1}^{N} N^k(x)\, I\{Z^k=i\}$ represents the number of observed deaths in the i^{th} sample at or before x-days.

The notation just presented is standard among authors, such as Gill[8], who have applied the theory of stochastic processes to survival analysis. It also will be useful to give the alternative notation for risk set sizes and numbers of deaths that was used by Mantel[3] in the development of the log-rank test.

In what follows, the subscript k will be used to index distinct, ordered, observed death (or event) times over all samples, with $k=1,2,\ldots,d^*$. Note that d^* will be the total number of observed deaths only if there are no tied

observed death times. Denote the set of ordered, distinct death times by $\{T_1, T_2, \ldots, T_{d^*}\}$ where $T_1<T_2<\ldots<T_{d^*}$. Then, $d_{ik} = dN_i(T_k)$ and $d_k = \sum_{i=0}^{r-1} dN_i(T_k)$ refer to the number of deaths occurring at T_k in the i^{th} sample and over all samples, respectively. The number of individuals in the risk set at death time T_k for the i^{th} sample and over all samples are $n_{ik} = Y_i(T_k)$ and $n_k = \sum_{i=0}^{r-1} Y_i(T_k)$, respectively.

Since all statistics to be defined are nonparametric, computations will be performed only over the interval in which risk set sizes are positive in at least two of the r samples. Thus, the last observed death time contributing information to the statistics is T_d, where

$$d \equiv \max[\, k: n_k > \max\{n_{ik},\ i=0,1,\ldots,r-1\}\,].$$

Clearly, $d \leq d^*$.

A2. Model. Denote the true survival distribution for the i^{th} sample by $S_i(t)$, which is simply the probability that an individual in sample i will survive from time 0 to time t. If we denote the cumulative hazard function in sample i by $\Lambda_i(t)$, it follows that $S_i(t) = \exp\{-\Lambda_i(t)\}$. Individuals in the i^{th} sample then have hazard function

$$\frac{d}{dt}\Lambda_i(t) \equiv \lambda_i(t) .$$

To help interpret the meaning of the hazard function, consider a small $\Delta\tau$. Then $\lambda_i(\tau)\Delta\tau$ is approximately the probability of death occurring in the interval τ to $\tau+\Delta\tau$, given survival to τ.

Let $C_i(t)$ denote the probability an individual in sample i is not censored before time t. Assuming statistical independence between the causes of death and censorship, it follows that the distribution of observation times in the i^{th} sample is given by $\pi_i(t) \equiv S_i(t)C_i(t)$.

Suppose $r \geq 2$. With the exception of one important special case, all test procedures to be discussed were developed to test the hypothesis that all r samples have equivalent survival distributions; that is,

$$H_o: \quad S_i(t) = S(t), \quad i=0,\ldots,r-1, \tag{1}$$

where S(t) is unspecified. The exception to this is that the two-sample linear rank tests to be discussed can be employed more generally to test the hypothesis

$$H_o': \quad \lambda_1(t) = \lambda_0(t)e^{\beta_0} \tag{2}$$

for some fixed β_0. Of course, H_o' reduces to H_o when $\beta_0=0$.

A3. Two-sample linear rank tests. As background, these test statistics are called rank tests because they depend on time only to the extent necessary to rank deaths and censored observations. A rank test statistic is invariant under any monotone transformation of the data because such a transformation does not alter the ranks. They are called linear because such statistics can be written as linear functions of the ranks.

A classic two-sample linear rank statistic to test H_o in censored survival data is the log-rank, proposed by Mantel[3]. Conditioning on risk set sizes, n_{0k} and n_{1k}, and on the number of events, d_k, occurring at T_k, Mantel proposed forming the difference between the observed and conditionally expected number of events in sample 1 at T_k. The log-rank statistic, as stated in (3), is then proportional to the sum of these differences when the sum is taken over event times:

$$\text{log-rank} \propto \sum_{k=1}^{d} \{d_{1k} - (\frac{n_{1k}}{n_{0k}+n_{1k}}) d_k\} \tag{3}$$

This statistic can be generalized to provide greater sensitivity to survival differences occurring over certain periods by employing a weight function $Q(T_k)$. One further generalization to test the more general

hypothesis H_o': $\lambda_1(t) = \lambda_0(t)e^{\beta_0}$ yields

$$Z^{(N)} \equiv \sum_{k=1}^{d} Q(T_k)\{ d_{1k} - \frac{n_{1k}e^{\beta_0}}{n_{0k}+n_{1k}e^{\beta_0}} d_k\} \tag{4}$$

Remembering that $Z^k \epsilon \{0,1\}$ in the two sample setting, equivalent formulations for $Z^{(N)}$, using Lebesgue-Stieltjes integrals, are given by:

$$Z^{(N)} = \sum_{k=1}^{N} \int_0^{\infty} Q(x)\{Z^k - \frac{\sum_{\ell=1}^{N} Z^\ell e^{\beta_0 Z^\ell} I\{X^\ell \geq x\}}{\sum_{\ell=1}^{N} e^{\beta_0 Z^\ell} I\{X^\ell \geq x\}}\} dN^k(x) \tag{5}$$

$$= \int_0^{\infty} Q(x) \{\frac{Y_0(x)\, Y_1(x)e^{\beta_0}}{Y_0(x)+Y_1(x)e^{\beta_0}}\} \{\frac{dN_1(x)}{Y_1(x)e^{\beta_0}} - \frac{dN_0(x)}{Y_0(x)}\} \tag{6}$$

$$= \int_0^{\infty} Q(x)\{\frac{Y_0(x)Y_1(x)e^{\beta_0}}{Y_0(x)+Y_1(x)e^{\beta_0}}\} d\{e^{-\beta_0}\hat{\Lambda}_1(x)-\hat{\Lambda}_0(x)\} \tag{7}$$

where (7) follows from (6) since the cumulative hazard estimator is

$$\hat{\Lambda}_i(x) \equiv \int_0^x \frac{dN_i(u)}{Y_i(u)} .$$

We can see from (7) that, when $\beta_0 = 0$, these two-sample linear rank statistics are simply a sum (integral) of weighted differences in hazard functions.

The variance, V, of $Z^{(N)}$ can be proposed heuristically by using weighted Bernoulli arguments in untied data, or in tied data with $\beta_0 = 0$, by hypergeometric distribution arguments of Mantel[3]. Employing the theory of stochastic processes, Gill[8] has verified that the statistic

$$\frac{Z^{(N)}}{V^{1/2}} = \frac{\sum_{k=1}^{d} Q(T_k)\{ d_{1k} - (\frac{n_{1k}e^{\beta_0}}{n_{0k}+n_{1k}e^{\beta_0}}) d_k\}}{\{\sum_{K=1}^{d} Q^2(T_k) \frac{n_{0k}n_{1k}e^{\beta_0}(n_{0k}+n_{1k}e^{\beta_0}-d_k)}{(n_{0k}+n_{1k}e^{\beta_0})^2 (n_{0k}+n_{1k}e^{\beta_0}-1)} d_k\}^{1/2}} \tag{8}$$

indeed has a standard normal distribution when $N\to\infty$, as long as Q satisfies some mild regularity conditions. To be precise, this result holds in untied data for any β_0 and holds when $\beta_0\equiv 0$ whether or not ties exist. However, when β_0 is a fixed non-zero constant and ties exist in the data, one must be more cautious. Formally, to study properties of statistics in tied data situations, one considers discrete time models. To amplify, many discrete time models exist which, as the discretization becomes finer, reduce to the proportional hazards model, $\lambda_1(t)=\lambda_0(t)e^{\beta}$. The statistics under these models in general will differ unless one is testing the hypothesis of equality; i.e., $H_o: \beta_0=0$. For example, one of these discrete time models, the so called "log odds" model of Cox[9], gives rise to a partial likelihood (equation 4.14 of Kalbfleisch and Prentice[10]) and a score statistic which agrees with our equation 8 (with $Q(t)\equiv 1$) only when $\beta_0\equiv 0$. Although differences between various discrete time statistics exist when testing a non-zero β_0, they are minor unless data are heavily tied. Thus, we suggest the use of the statistic in equation (8) due to its simplicity, ease of computation, and intuitive appeal. Still, one should be cautious if $\beta_0\neq 0$ and data are heavily tied.

Observe in equation (8) that the weighted observed number of deaths in sample 1 is given by

$$\sum_{k=1}^{d} Q(T_k)\, d_{1k}$$

while the weighted expected number of deaths in that sample 1 is

$$\sum_{k=1}^{d} Q(T_k)\left(\frac{n_{1k}e^{\beta_0}}{n_{0k}+n_{1k}e^{\beta_0}}\right) d_k$$

As mentioned earlier, the weight function Q enables one to obtain particular sensitivity to survival differences occurring at specific points in time. The following table indicates those weight functions which we are making available and the name of the corresponding test statistic. $\hat{S}$ is simply the left-continuous Kaplan-Meier[12] survival estimator in the pooled sample.

Q(x)(weight function)	Test Statistic
$Y_0(x)+Y_1(x)$	Gehan-Wilcoxon[1]
$[\hat{S}(x)]^\rho \ [1-\hat{S}(x)]^\gamma$	$G^{\rho,\gamma}$ class
1	log-rank[3] ($G^{\rho,\gamma}$, $\rho=0$, $\gamma=0$)
$\hat{S}(x)$	Peto-Peto Wilcoxon[4] ($G^{\rho,\gamma}$, $\rho=1$, $\gamma=0$)
$[\hat{S}(x)]^\rho$	Harrington-Fleming G^ρ[2] ($G^{\rho,\gamma}$, $\gamma=0$)

Briefly, relative to the log rank test, the Gehan-Wilcoxon and Peto-Peto Wilcoxon provide greater sensitivity to survival differences occurring earlier in time since $(Y_0(x)+Y_1(x))$ and $\hat{S}(x)$ are decreasing weight functions. The Harrington-Fleming G^ρ class includes the log-rank ($\rho=0$) and Peto-Peto Wilcoxon ($\rho=1$) as special cases and provides greater sensitivity to early occurring differences the larger one chooses ρ. For the situation in which $\beta_0=0$, Harrington and Fleming[2] have found the type of departures from H_o that each of the G^ρ test produces is fully efficient in detecting. Obviously, the $G^{\rho,\gamma}$ family provides considerable versatility to the user. Sensitivity to early occurring differences is obtained by taking $\rho>0$, $\gamma\approx0$, to middle differences by taking $\rho\approx1\approx\gamma$, and to late occurring differences by taking $\rho\approx0$, $\gamma>0$.

A4. r-sample linear rank tests of equality of survival ($e^{\beta_0}\equiv1$)

Recall in two samples from equation (8), with $e^{\beta_0}\equiv1$, $\frac{Z^{(N)}}{V^{1/2}} \sim N(0,1)$ with formulation

$$\frac{\sum_{k=1}^{d} Q(T_k)\left\{d_{1k}-\left(\frac{n_{1k}}{n_k}\right) d_k\right\}}{\left\{\sum_{k=1}^{d} Q^2(T_k)\ \frac{n_{0k}n_{1k}\ (n_k-d_k)}{(n_k)^2\ (n_k-1)}\ d_k\right\}^{1/2}}$$

This statistic can be generalized to r-samples as $\underset{\sim}{Z}^{(N)'}\ \underset{\sim}{V}^{(N)^{-1}}\ \underset{\sim}{Z}^{(N)} \sim \chi^2_{r-1}$ where

$$\left(\underset{\sim}{Z}^{(N)}\right)_i = \sum_{k=1}^{d} Q(T_k)\left\{d_{ik} - \frac{n_{ik}}{n_k}\, d_k\right\}$$

and

$$\left(\underset{\sim}{V}^{(N)}\right)_{i\ell} = \sum_{k=1}^{d} Q^2(T_k)\, \frac{n_{ik}}{n_k}\left(\delta_{i\ell} - \frac{n_{\ell k}}{n_k}\right) d_k\left(\frac{n_k - d_k}{n_k - 1}\right)$$

where

$\delta_{i\ell}=1$ if $i=\ell$ and 0 if $i\neq\ell$.

The weighted, observed number of deaths in the i^{th} sample is $\sum_{k=1}^{d} Q(T_k)\, d_{ik}$, with corresponding weighted, expected number of deaths

$$\sum_{k=1}^{d} Q(T_k)\, \frac{n_{ik}}{n_k}\, d_k \quad \text{for } i=0,1,\ldots\ r-1.$$

For

$$Q(x) = \left[\hat{S}(x)\right]^{\rho}\left[1-\hat{S}(x)\right]^{\gamma} \tag{9}$$

we have the $G^{\rho,\gamma}$ class, while for $Q(x) = \sum_{i=0}^{r-1} Y_i(x)$, we have the Breslow-Gehan-Wilcoxon[11] test statistic.

A5. <u>One-sample linear rank, goodness-of-fit tests ($e^{\beta_0}=1$)</u>. The class of one-sample goodness-of-fit tests[2] which we present can be obtained from equation (6), with $\beta_0=0$ and Q(x) as defined in (9), by letting $N_1\to\infty$. The hypothesis to be tested is that the true underlying survival function, S, is equal to some specified S_0. In the one sample problem, note that $N\equiv N_0$.

The statistic's numerator can be shown to be

$$Z^{(N)}_{\rho,\gamma} = \sum_{k=1}^{N} \int_0^{X^k} \left\{S_0(x)\right\}^{\rho}\left\{1-S_0(x)\right\}^{\gamma} d\Lambda_0(x)$$

$$- \sum_{k=1}^{N} \left\{S_0(X^k)\right\}^{\rho}\left\{1-S_0(X^k)\right\}^{\gamma} \Delta^k\ .$$

Observe that the statistic continues to be the difference between the sum of weighted expected and weighted observed numbers of deaths.

It can be shown that $N^{-1/2} Z_{\rho,\gamma}^{(N)} \sim N(0,\sigma^2)$ as $N\to\infty$

with $\sigma^2 = \int_0^\infty \{S_0(x)\}^{2\rho} \{1-S_0(x)\}^{2\gamma} \pi_0(x) d\Lambda_0(x)$. The variance σ^2 is consistently estimated by $N^{-1}V = N^{-1} \sum_{k=1}^{N} \int_0^{X^k} \{S_0(x)\}^{2\rho}\{1-S_0(x)\}^{2\gamma}\, d\Lambda_0(x)$.

Then $\dfrac{Z_{\rho,\gamma}^{(N)}}{V^{1/2}} \sim N(0,1)$.

For example, with $\rho>0$, $\gamma=0$, we have

$$\frac{Z_{\rho,0}^{(N)}}{V^{1/2}} = \frac{\sum_{k=1}^{N} \rho^{-1}[1-\{S_0(X^k)\}^{\rho}] - \sum_{k=1}^{N} \Delta^k \{S_0(X^k)\}^{\rho}}{\left\{\sum_{k=1}^{N} (2\rho)^{-1}[1-\{S_0(X^k)\}^{2\rho}]\right\}^{1/2}} .$$

By setting $\rho=0$, $\gamma=0$, we obtain the one-sample version of the log-rank test, which is given by

$$\frac{Z_{0,0}^{(N)}}{V^{1/2}} = \frac{\sum_{k=1}^{N} [-\ell n\, S_0(X^k)] - \sum_{k=1}^{N} \Delta^k}{\left\{\sum_{k=1}^{N} [-\ell n\, S_0(X^k)]\right\}^{1/2}} .$$

Note that to calculate these one-sample statistics, one need only specify (X^k, Δ^k) and $S_0(X^k)$ for the k^{th} individual; k=1, ..., N. As stated above, the one-sample $G^{\rho,\gamma}$ test statistics are only available for $\rho\geq 0$, $\gamma=0,1,2$.

A6. Two-sample Non-linear Rank Tests. Through this SAS procedure, one can compute two-sample supremum-type non-linear rank tests of equality of survival distributions. These tests are based upon the "Kappa Rho" (κ^ρ) class of statistics[6] and the generalized Smirnov[5] statistic.

To provide motivation for the κ^ρ class of statistics note that in biological problems and in many other areas of application, the proportional hazards model frequently is thought to adequately represent the relationship of a covariate to a continuous endpoint. As a result, the Cox partial likelihood score statistic and more specifically the logrank statistic, G^0, provide the standard against which other censored data rank statistics must be compared. The Renyi-type statistic κ^0 is essentially a supremum version of the logrank statistic. κ^0 provides one natural way to obtain a test procedure which nearly maintains the sensitivity of G^0 against proportional hazards alternatives, yet which is more powerful than G^0 when the hazard ratio is clearly non-constant. Simulation studies do confirm the high power of κ^0 against proportional hazards alternatives and confirm that κ^0 is more versatile than G^0 across several distinctly different configurations in uncensored or lightly censored data[6]. Similarly, it is apparent that κ^ρ provides a supremum version of the linear rank statistic G^ρ, with κ^ρ being more versatile in uncensored or lightly censured data.

To formulate the κ^ρ statistics, choose a value of $\rho \geq 0$. Let $\hat{S}_i(t)$ and $\hat{C}_i(t)$ denote estimates of $S_i(t)$ and $C_i(t)$. Then

$$\kappa^\rho = (\tilde{\sigma}^2)^{-1/2} \sup_{t \geq 0} \kappa^\rho(t)$$

where

$$\kappa^\rho(t) = \int_0^t \frac{1}{2}\{[\hat{S}_0(x)]^{\rho+1} + [\hat{S}_1(x)]^{\rho+1}\}\left\{\frac{N_0\hat{C}_0(x)N_1\hat{C}_1(x)}{N_0\hat{C}_0(x)+N_1\hat{C}_1(x)}\right\}^{1/2} * I\{N_0(x)N_1(x)>0\}\left\{\frac{dN_0(x)}{Y_0(x)} - \frac{dN_1(x)}{Y_1(x)}\right\}$$

and where

$$\tilde{\sigma}^2 = \int_0^\infty \frac{1}{2}\{[\hat{S}_0(x)]^{2\rho+1} + [\hat{S}_1(x)]^{2\rho+1}\} I\{N_0(x)N_1(x)>0\}\left\{\frac{dN_0(x)+dN_1(x)}{N_0(x)+N_1(x)}\right\}.$$

Note that for $\alpha=\rho+1$, κ^{ρ} is identical to the K^{α} statistic of Fleming and Harrington[7] except for the formulation of $\tilde{\sigma}^2$. κ^{ρ} is superior to K^{α} in the sense that κ^{ρ} has appropriate size even in small and moderate sample size applications.

Another supremum-type test statistic is the generalized Smirnov (GS) test. It is formulated as follows:

$$GS = \sup_{t} Y_{N_0,N_1}(t)$$

$$\text{where } Y_{N_0,N_1}(t) = \frac{1}{2}\{\hat{S}_0(t^+)+\hat{S}_1(t^+)\} \int_0^t \left\{\frac{N_0\hat{C}_0(x)\ N_1\hat{C}_1(x)}{N_0\hat{C}_0(x)+N_1\hat{C}_1(x)}\right\}^{1/2}$$

$$* I\{Y_0(x)\ Y_1(x)>0\} \left\{\frac{dN_1(x)}{Y_1(x)} - \frac{dN_0(x)}{Y_0(x)}\right\} ,$$

$$\text{where } \hat{S}_i(t^+) \equiv \lim_{u \downarrow t} \hat{S}_i(u) .$$

The GS statistic is a versatile test statistic with sensitivity to any survival differences which are large at some point in time, independent of the type of differences existing elsewhere. The corresponding test produced is especially sensitive to departures from H_o in which the two survival distributions exhibit a substantial difference in their middle range, but possibly have this difference disappear when hazard functions cross.

B. SURVDIFF Statement Specification

PROC SURVDIFF options;

Options:

DATA=data_set_name

Specifies the name of the data set to be used. If omitted, it uses the last one created.

GW

Indicates the Gehan-Wilcoxon statistic. This option is appropriate for 1-, 2- and r>2-sample problems.

GS

Requests generalized Smirnov test. Option appropriate for two-sample problems only.

SAMPLES=r

Specifies the number of samples. Use 1, 2, and R for 1, 2, and $r \geq 3$ samples, respectively. If omitted, data will be scanned for number of samples. If used, it speeds up processing. There are two methods of specifying the number of samples:

i) If the "SAMPLES=r" option is not used, the procedure scans the data set or by-group to determine the number of levels of the CLASS variable and performs only those requested tests appropriate for the number of samples.

ii) If the "SAMPLES=r" option is used, the procedure requires that the data set or each by-group meet the following criteria:

a. If SAMPLES=1 is specified, SOFT= must be coded and the CLASS statement omitted.

b. If SAMPLES=2 is specified, tests are performed only on the data set or by-groups which have 2 levels of the variable in the CLASS statement.

c. If SAMPLES=R is specified, tests are performed only on the data set or by-groups which have ≥ 3 levels of the variable in the CLASS statement.

VARNAMES [keyword$_i$=key_word_variable_name$_i$];

Keyword

TIME=variable_name_of_observation_time

This numeric variable contains the time in days (actually any time units are permissible) from time 0 to the event of interest or censoring. If not specified, the variable, TIME, appropriately defined is assumed to exist on the data set.

EVENT=variable_name_of_event_indicator

EVENT indicates whether the observation time is an event (death) time or time of censorship where 1=censor, 2=event. If not specified, the variable EVENT appropriately defined is assumed to exist on the data set.

SOFT=variable_name_of_hypothesized_survival

SOFT indicates the expected survival probability, $S_0(t)$, used in the one-sample, linear rank tests. It is specific to the individual and to the individual's observation time. SOFT is required and only appropriate for one-sample problems.

CLASS class_variable_name;

The class variable name defines the groups to be compared. It may be a numeric or character variable. If character, the maximum length is sixteen characters.* Omit the CLASS statement for one-sample problems.

*If your character variable is longer than 16, you will be in trouble. They (GAS SAS) are NOT responsible.

GRHOGAMMA $([\{\rho_1\}]) * ([\{\gamma_1\}]) \cdot \cdot \cdot ([\{\rho_K\}]) * ([\{\gamma_K\}])$;

The notation $([\{\rho_i\}])*([\{\gamma_i\}])$ refers to a set of ρ's and γ's to be used. The limit on the number of $\rho*\gamma$ combinations is 100. These options are appropriate for 1-, 2-, and r $\geq$3 sample problems.

Examples: GRHOGAMMA 0*0 1*0; would result in a log-rank and Peto-Peto Wilcoxon test, respectively, or their analogs for 1-, 2-, and r$\geq$3 sample situations.

Coding GRHOGAMMA (0 1) * (1 2 3); is equivalent to coding GRHOGAMMA 0*1 0*2 0*3 1*1 1*2 1*3; At least one ρ ($0 \leq \rho \leq 32767$) and γ ($0 \leq \gamma \leq 32767$) combination must be specified with the GRHOGAMMA statement.

BETA β_1 β_2 $\cdot \cdot \cdot$ β_k; or BETA β_1 TO β_2 BY increment;

BETA indicates the β term for testing departures from a proportional hazards model with hazard ratio exp(β), as described above. The default is β=0, which corresponds to testing the equality of the hazard and thus equality of survival. Non-zero values for BETA are appropriate for two-sample problems only. A maximum of 100 BETA's may be specified where $-32767 \leq BETA \leq 32767$. The increment is any positive number such that $0 < increment \leq 32767$.

KAPPARHO ρ_1 ρ_2 • • • ρ_k;

KAPPARHO specifies the ρ_i values to be used in the KAPPARHO statistic. The maximum number of rho values that can be specified is 100, where $-1 \leq \rho_i \leq 32767$. This option is appropriate for the two-sample situation only.

BY by_variable(s);

As with other procedures, analysis will be done separately for each level of the by-variable provided the data are sorted accordingly.

C. Testing One-Sided Hypotheses

Only for the 2-sample situation can a one-sided test be computed.

For κ^ρ and generalized Smirnov statistics, one-sided p-values are given where, under the alternative hypothesis, the second sample in sort order of the class variable is the one assumed to have longer (better) survival.

For the $G^{\rho,\gamma}$ class of statistics, specification of the level of the class variable must be done in conjunction with specification of β. Let $\lambda_1(t)$ be the hazard for sample 1 (first in sort order of the class variable) and $\lambda_2(t)$ the hazard for sample 2. If $\lambda_2(t)<\lambda_1(t)$ the survival in sample 2 is longer (better) than sample 1; i.e. smaller hazard implies better survival. The formulation in this procedure is $\lambda_2(t)=\lambda_1(t)e^\beta$. Thus, if β is negative, $e^\beta<1$ and it follows that the second sample in sort order of the class variable is hypothesized to have longer survival than the first sample in sort order. For β positive, the first sample in sort order is hypothesized to have longer survival than the second. To be consistent with the supremum statistics, when setting $\beta \neq 0$, it is necessary to specify the appropriate negative β and define the class variable so that the second in sort order has the longer hypothesized survival. A note appears on the output to aid in the interpretation for one-sided tests.

For all the linear rank statistics, only two-tail P-values are printed, but one-tail P-values can be easily calculated. Suppose, for example, $\beta \neq 0$ and the alternative of interest is in the direction of $\beta=0$. If in the sample with hypothesized longer survival, the sum of the weighted observed events is greater than the sum of the weighted expected, then the one-tail P-value is one-half of the two-tail P-value. If, in the sample with hypothesized longer survival, this sum of weighted observed events is less than the sum of the weighted expected, then the one-tail P-value is 1-(.5*two-tail P-value).

As already mentioned, one-tail P-values are printed for the κ^{ρ} and the generalized Smirnov statistics. Note that for the supremum statistics, the one-sided P-value is not simply 1/2 of the two-sided P-value.

D. General Comments

There are no default test statistics. Desired test statistics must be specified.

Only one of each of the statements is permitted.

Two-sided P-values are always printed. For the generalized Smirnov statistic, two-tail P-values above 0.80 are difficult to compute precisely (see ref. 5) and thus are denoted ">0.80".

Label and format capabilities are available only for the CLASS variable.

Since the availability of test statistics and options is specific to the number of samples (groups) being compared, we prepared the following table.

Test Statistics Available

No. of Samples (r)	$G^{\rho,\gamma}$ (GRHOGAMMA)	Gehan-Wilcoxon (GW)	K^{ρ} (KAPPARHO)	Generalized Smirnov (GS)
1	Yes ($\rho \geq 0$, γ=0,1,2)	Yes	No	No
2 (BETA=0)	Yes ($\rho \geq 0$, $\gamma \geq 0$)	Yes	Yes	Yes
2 (BETA≠0)	Yes ($\rho \geq 0$, $\gamma \geq 0$)	Yes	No	No
$\geq$3 (BETA=0)	Yes ($\rho \geq 0$, $\gamma \geq 0$)	Yes	No	No

Statement and Option Specifications†

No. of Samples (r)	PROC options				CLASS stmt	VARNAMES stmt			GRHOGAMMA stmt	KAPPARHO stmt	BETA stmt
	GW	GS	SAMPLES=	DATA=		EVENT=	TIME=	SOFT=			
1	NA	NA	omit, 1	OPT	NA	OPT*	OPT*	REQ	REQ	NA	NA
2(β=0)	OPT	OPT	omit, 2	OPT	REQ	OPT*	OPT*	NA	OPT	OPT	OPT (default =0)
2(β≠0)	OPT	NA	omit, 2	OPT	REQ	OPT*	OPT*	NA	OPT	NA	REQ
$\geq$3(β=0)	OPT	NA	omit, R	OPT	REQ	OPT*	OPT*	NA	OPT	NA	NA

† NA not applicable or appropriate. If specified an error will result and processing will stop.

OPT optional. Note that at least one test statistic must be specified.

REQ required.

* If not specified, appropriately defined variables with variable names EVENT and TIME respectively are assumed to exist on the data set.

E. Example

The example presented below is taken from the study of Lininger et. al.[13] and is used as an example by Fleming et. al.[14] to sequentially test the comparability of survival in the two regimens. The results presented below correspond to Fleming et. al.'s[14] first analysis date of 9/12/77. The primary goal was to evaluate the survival experience among patients having extensive stage small cell lung cancer, comparing two regimens of chemotherapy. In regimen A, the "experimental" treatment, patients received cyclophosphamide (CTX), vincristine (VCR), VP-16, and cis-platin (CDDP) alternating with adriamycin (ADR) and imidazole carboxamide (DTIC). Regimen B, the "standard" treatment, consisted of ADR, VCR, VP-16 and CDDP alternating with CTX and DTIC.

SMALL CELL LUNG CANCER SURVIVAL DATA.
STATUS: 1=CENSORED 2=DEATH

REGIMEN=_B.STND			REGIMEN=A_.EXPT		
OBS	DAYS	STATUS	OBS	DAYS	STATUS
1	19	1	18	8	1
2	119	1	19	47	1
3	136	1	20	62	1
4	216	1	21	87	1
5	312	1	22	98	1
6	398	1	23	155	1
7	9	2	24	166	1
8	10	2	25	187	1
9	99	2	26	223	1
10	122	2	27	335	1
11	148	2	28	373	1
12	228	2	29	383	1
13	233	2	30	395	1
14	280	2	31	402	1
15	282	2	32	488	1
16	375	2	33	22	2
17	420	2	34	142	2
			35	171	2
			36	635	2

```
PROC SURVDIFF  GW    GS ;
   GRHOGAMMA   (0 1)*0  ;
   KAPPARHO    0        ;
   CLASS       REGIMEN  ;
   VARNAMES    TIME=DAYS  EVENT=STATUS ;
TITLE2 EXAMPLE OF LOG-RANK, PETO-PETO-WILCOXON, GEHAN-WILCOXON, ;
TITLE3 KAPPA-RHO & GENERALIZED SMIRNOV STATISTICS.              ;
```

A

EXAMPLE OF LOG-RANK, PETO-PETO-WILCOXON, GEHAN-WILCOXON, KAPPA-RHO & GENERALIZED SMIRNOV STATISTICS.

G RHO GAMMA TEST (1) (LOGRANK)

FOR VARIABLES: TIME=DAYS EVENT=STATUS (2)

(3) BETA= 0.0000 RHO= 0.00 GAMMA= 0.00

REGIMEN (4)	N (5)	OBS. (6)	NUMBER OF EVENTS (7) O=SUM OF WEIGHTED OBSERVED	(8) E=SUM OF WEIGHTED EXPECTED	(O-E)**2/E (9)
_B.STND	17	11	11.00	6.38	3.35
A_.EXPT	19	4	3.00	7.62	2.81
TOTAL	36	15	14.00	14.00	6.16

CHI SQUARE= 6.23 (10) DF= 1 (11) TWO-TAILED P= 0.0126 (12)

**** OBS. DELETED DUE TO: MISSING VALUES= 0 (13) ****
**** INVALID DATA= 0 ****

G RHO GAMMA TEST (PETO-PETO WILCOXON)

FOR VARIABLES: TIME=DAYS EVENT=STATUS

BETA= 0.0000 RHO= 1.00 GAMMA= 0.00

REGIMEN	N	OBS.	NUMBER OF EVENTS O=SUM OF WEIGHTED OBSERVED	E=SUM OF WEIGHTED EXPECTED	(O-E)**2/E
_B.STND	17	11	8.21	4.97	2.10
A_.EXPT	19	4	2.56	5.79	1.80
TOTAL	36	15	10.76	10.76	3.90

CHI SQUARE= 4.89 DF= 1 TWO-TAILED P= 0.0271

**** OBS. DELETED DUE TO: MISSING VALUES= 0 ****
**** INVALID DATA= 0 ****

B

EXAMPLE OF LOG-RANK, PETO-PETO-WILCOXON, GEHAN-WILCOXON, KAPPA-RHO & GENERALIZED SMIRNOV STATISTICS.

GEHAN-WILCOXON TEST

FOR VARIABLES: TIME=DAYS EVENT=STATUS

BETA= 0.0000

REGIMEN	N	OBS.	NUMBER OF EVENTS O=SUM OF WEIGHTED OBSERVED	E=SUM OF WEIGHTED EXPECTED	(O-E)**2/E
_B.STND	17	11	5.78	3.67	1.22
A_.EXPT	19	4	2.06	4.17	1.07
TOTAL	36	15	7.83	7.83	2.29

CHI SQUARE= 3.35 DF= 1 TWO-TAILED P= 0.0671

**** OBS. DELETED DUE TO: MISSING VALUES= 0 ****
**** INVALID DATA= 0 ****

KAPPA-RHO TEST

FOR VARIABLES: TIME=DAYS EVENT=STATUS

(14) RHO= 0.00

REGIMEN	N	OBSERVED EVENTS
_B.STND	17	11
A_.EXPT	19	4
TOTAL	36	15

ONE-TAILED P= 0.0029 TWO-TAILED P= 0.0059

--NOTE:
FOR THE 1-SIDED TEST, SAMPLE A_.EXPT
IS HYPOTHESIZED TO HAVE LONGER (BETTER)
SURVIVAL THAN SAMPLE _B.STND

**** OBS. DELETED DUE TO: ****
**** MISSING VALUES= 0 ****
**** INVALID DATA= 0 ****

C

EXAMPLE OF LOG-RANK, PETO-PETO-WILCOXON, GEHAN-WILCOXON, KAPPA-RHO & GENERALIZED SMIRNOV STATISTICS.

GENERALIZED SMIRNOV

FOR VARIABLES: TIME=DAYS EVENT=STATUS

REGIMEN	N	OBSERVED EVENTS
_B.STND	17	11
A_.EXPT	19	4
TOTAL	36	15

(15) V= 1.7554 A= 1.7554 R= 0.5557
TIME(V)= 420 TIME(A)= 420

ONE-TAILED P= 0.0016 TWO-TAILED P= 0.0032

--NOTE:
FOR THE 1-SIDED TEST, SAMPLE A_.EXPT
IS HYPOTHESIZED TO HAVE LONGER (BETTER)
SURVIVAL THAN SAMPLE _B.STND

**** OBS. DELETED DUE TO: ****
**** MISSING VALUES= 0 ****
**** INVALID DATA= 0 ****

```
PROC SURVDIFF ;
   GRHOGAMMA   0*0     ;
   CLASS       REGIMEN ;
   VARNAMES    TIME=DAYS EVENT=STATUS ;
   BETA        -0.693  ;
TITLE2 EXAMPLE OF "LOG-RANK" STATISTIC FOR TESTING A SPECIFIED  ;
TITLE3 DIFFERENCE IN HAZARDS. THE NULL HYPOTHESIS IS :          ;
TITLE4 HAZARD FOR A_.EXPT=HAZARD FOR _B.STND*(EXP(BETA)). WITH  ;
TITLE5 BETA=-0.693, EXP(BETA)=0.5 . PROPORTIONAL HAZARDS ASSUMED.;
```

D

EXAMPLE OF "LOG-RANK" STATISTIC FOR TESTING A SPECIFIED DIFFERENCE IN HAZARDS. THE NULL HYPOTHESIS IS : HAZARD FOR A_.EXPT=HAZARD FOR _B.STND*(EXP(BETA)), WITH BETA=-0.693, EXP(BETA)=0.5 . PROPORTIONAL HAZARDS ASSUMED.

G RHO GAMMA TEST

FOR VARIABLES: TIME=DAYS EVENT=STATUS

BETA= -0.6930 RHO= 0.00 GAMMA= 0.00

REGIMEN	N	OBS.	NUMBER OF EVENTS O=SUM OF WEIGHTED OBSERVED	E=SUM OF WEIGHTED EXPECTED	(O-E)**2/E
_B.STND	17	11	11.00	8.74	0.59
A_.EXPT	19	4	3.00	5.26	0.97
TOTAL	36	15	14.00	14.00	1.56

CHI SQUARE= 1.58 DF= 1 TWO-TAILED P= 0.2089 (16)

--NOTE: UNDER THIS NULL HYPOTHESIS, SAMPLE A_.EXPT
HAS LONGER (BETTER) SURVIVAL
THAN SAMPLE _B.STND

**** OBS. DELETED DUE TO: MISSING VALUES= 0 ****
**** INVALID DATA= 0 ****

(1) Test being specified.

(2) Variable names which reflect time and event status, respectively.

(3) The beta, rho, gamma specifications. Note default beta is zero.

(4) The class variable and sample names.

Note: to get the samples into the desired sort order for subsequent one-sided tests we defined the standard regimen B as '_B.STND' and the experimental regimen A the value 'A_.EXPT' so that the sample with the hypothesized better survival would appear second in SAS sort order.

(5) Number of valid observations in sample.

(6) Sum of observed number of events in sample.

(7) Sum of weighted observed number of events.

(8) Sum of weighted expected number of events.

Note: When rho=gamma=0 the weights are unity. When rho>0 or gamma>0 are specified, the weights are, in general, values less than unity; hence, in this situation the totals for (7) and (8) are less than the total (6).

Also, the linear rank statistics are non-parametric procedures. As such, in the 2-sample problem, for example, events occurring in one sample after the longest event or censorship time in the other sample are ignored. In this example, note that the total for (6) is 15, while for (7) and (8) it is 14. The death at 635 days in the regimen B_.EXPT is in essence ignored.

(9) As an exploratory tool, this displays the samples for which differences between weighted observed and expected number of events stand out. The total conservatively estimates the correct chi-square value printed at (10).

(11) Degrees of freedom.

(12) Two-tail P-value associated with test of H_0 which in this case is the test of equality of survival curves.

(13) Observations will be deleted and noted here if any of the following conditions hold:

- missing values: if the time-variable, event-indicator-variable, or SOFT-variable as required are missing.
- invalid data: the time-variable is negative or the status-variable is other than a 1 or 2. SOFT, if used, must satisfy $0 \leq SOFT \leq 1$.

(14) The rho value specified in the KAPPARHO statement.

(15) V is the maximum of the one-sided, standardized differences. TIME(V) is the time in days when V was observed. A is the maximum of the two-sided, standardized differences. TIME(A) is the time in days when A was observed. R is the average survival over the two samples at the largest time which satisfies $N_0(t)*N_1(t) > 0$. In this example, the time would be 420 days.

(16) This two-tail P-value tests whether or not the ratio of the hazards for regimen A_.EXPT over regimen _B.STND differs from 0.5 (i.e., $e^{-0.693}=0.5$), assuming a constant hazard ratio over all times. To obtain a one-sided P-value associated with $H_o: \lambda_A(t)/\lambda_B(t) \leq 0.5$ versus $H_A: \lambda_A(t)/\lambda_B(t) > 0.5$ where $\lambda_A(t)$ and $\lambda_B(t)$ refer to the population hazard for regimen A and B respectively, we focus on the sum of the weighted observed and the sum of the weighted expected numbers for the A_.EXPT regimen. We see a value of 3.00 for the sum of the weighted observed number of deaths and 5.26 for the sum of the weighted expected number of deaths. The one-sided P-value is hence $1-[1/2 \times 0.2089] = 0.89555$. Had the sum of the weighted observed been greater than the sum of the weighted expected, the one-tail P-value would have been $[1/2 \times 0.2089] = 0.10445$.

In summary, the hazard ratio as described above is not detectably different from 0.5 in the two-sided context (P=0.2089). Also, the hazard ratio is not detectably greater than 0.5 in the one-sided context (P=0.89555).

Acknowledgements

We would like to thank Mr. Jon Kosanke for his work in writing the parsor and Ms. Marilyn Ness for her able assistance in typing the manuscript.

References

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

(2) Harrington, D.P., Fleming, T.R. (1982). A class of rank test procedures for censored survival data. Biometrika, 69, 553-566.

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

(4) Peto, R., Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). Journal of the Royal Statistical Society, Series A, 135, 185-206.

(5) Fleming, T.R., O'Fallon, J.R., O'Brien, P.C. and Harrington, D.P. (1980). Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right censored data. Biometrics, 36, 607-625.

(6) Fleming, T.R., Harrington, D.P. (1984). Supremum Versions of the Log Rank and Generalized Wilcoxon Statistics. Mayo Clinic Technical Report #24. (submitted for publication)

(7) Fleming, T.R., Harrington, D.P. (1981). A class of hypothesis tests for one and two sample censored survival data. Communications in Statistics, A10(8), 763-794.

(8) Gill, R.D. (1980). Censoring and Stochastic Integrals. Mathematical Centre Tracts 124, Mathematische Centre, Amsterdam.

(9) Cox, D.R. (1972). Regression Models and Life Tables. Journal of the Royal Statistical Society, Series B, 34:187-220.

(10) Kalbfleisch, J.D., Prentice, R.L. (1980). The Statistical Analysis of Failure Time Data, Wiley.

(11) Breslow, N.E. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika, 57, 579-594.

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

(13) Lininger, T.R., Fleming, T.R., Eagan, R.T. (1981). Evaluation of alternating chemotherapy and sites and extent of disease in extensive small cell lung cancer. Cancer, 48:2147-2153.

(14) Fleming, T.R., Harrington, D.P., O'Brien, P.C. (1984). Designs for group sequential tests. Controlled Clinical Trials, (in press).