LIFETEST+ODS+IML=Stratified Log Rank Tests
Ann Olmsted, CV Therapeutics, Palo Alto, CA

ABSTRACT
When analyzing right-censored survival data, equality of survival curves across treatment groups is often tested using log rank tests (Peto and Peto 1972), aka Mantel-Haenszel tests. For instance, in a cancer trial, the primary response variable might be time to disease progression and the treatment groups might be chemotherapy regimens. To increase power and protect against baseline imbalances, you could stratify by a variable with known prognostic value, such as disease stage. But in versions 8 and earlier, proc LIFETEST performs unstratified log rank tests only. Sorting the input dataset by stratum and adding a "by stratum;" statement yields the information you need to hand-compute the stratified log rank statistic. By feeding ODS output datasets to SAS/IML, you can easily automate these computations and obtain exactly the same stratified log rank test p-values calculated by SPSS, Stata, and BMDP. Stratified generalized Wilcoxon tests and 1 df trend tests are a bonus.

INTRODUCTION
Using Miller’s notation (Miller 1981, p. 107), assume the survival times for the $i^{th}$ treatment group are a random sample from a distribution $F_i$, $i=1,...,K$, assume that survival and censoring times are independent, and assume that the censoring times for the $i^{th}$ treatment group are a random sample from a distribution $G_i$, $i=1,...,K$. We observe $(X_{ij}, \delta_{ij}), i=1,...,K$, $j=1,...,n_i$, where $X_{ij}$ is the survival time for the $j^{th}$ individual in the $i^{th}$ treatment group and $\delta_{ij}$ indicates whether the observed time is exact or right-censored: $\delta_{ij} = I(T_{ij} \leq C_{ij})$.

We want to test either $H_0: F_1 = \cdots = F_K$ or $H^*_0: F_1 = \cdots = F_K$ and $G_1 = \cdots = G_K$.

That is, we want to test equality of the survival distributions, with or without assuming equality of the censoring distributions. For some clinical trials, censoring patterns are clearly equal: e.g., when patient enrollment begins on a certain date, patients are randomly assigned to treatments, patient follow-up ends on a certain date, and no patients are lost to follow-up before that date. For other clinical trials, such as those in which patients may be lost to follow-up because of adverse effects, censoring patterns can vary.

THE LOG RANK TEST
Denote the pooled sample, disregarding group membership, as $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$.

When the censoring patterns are equal, then conditional on the pooled sample, every possible division into $K$ groups of size $n_1, \ldots, n_K$ is equally likely. Given an appropriate test statistic, say $T$, with expectation $0$ under $H^*_0$, we can find the permutation covariance matrix of $T$ under $H^*_0$, $\Sigma_0$, and hope to show that $T'\Sigma_0^{-1}T$, where $\Sigma_0^{-1}$ is a generalized inverse of $\Sigma_0$, has a chi-squared limiting distribution under $H^*_0$. This in fact is the basis for one version of the generalized Wilcoxon test, the Gehan test discussed below. But when the censoring patterns cannot be assumed equal, a somewhat different approach is needed. The approach implemented in proc LIFETEST is to condition on the distinct exact times, the number of individuals at risk just before each exact time, and the number of deaths at each exact time:

\[
\begin{array}{cccc}
1 & 2 & \cdots & K \\
d_{11} & d_{12} & \cdots & d_{1K} \\
d_{21} & d_{22} & \cdots & d_{2K} \\
& & \vdots & \vdots \\
d_{K1} & d_{K2} & \cdots & d_{KK} \\
\end{array}
\]

\[
\begin{array}{cccc}
a_{11} & a_{12} & \cdots & a_{1K} \\
a_{21} & a_{22} & \cdots & a_{2K} \\
& & \vdots & \vdots \\
a_{K1} & a_{K2} & \cdots & a_{KK} \\
\end{array}
\]

Under $H_0$, the number of deaths in each group at the $l^{th}$ distinct exact time, $D_l = (d_{l1}, \ldots, d_{lK})’$, has a multivariate hypergeometric distribution, with expectation

\[
\left( \frac{d_{l1}r_{1l}}{r_1 + \cdots + r_K}, \ldots, \frac{d_{lK}r_{Kl}}{r_1 + \cdots + r_K} \right),
\]

where $d_{li}$ is the number of deaths at the $l^{th}$ exact time and $r_{ij}$ is the number at risk in group $j$ just before that time. The expression for the conditional covariance matrix of $D_l$ under $H_0$, $\Sigma_0$, is more complicated.

The covariance matrix of $T$ under $H_0$ is conditionally equal and unconditionally approximately equal to the sum of the individual covariance matrices: $\Sigma_0(T) = \sum_i \Sigma_{0_i}$. Although the individual tables of counts at each exact time are not independent, it can be shown that $T’ \left( \sum_i \Sigma_{0_i} \right)^{-1} T$ has a chi-squared limiting distribution under $H_0$, just as it does when $T$ is a genuine Mantel-Haenszel statistic. This is the basis for the log rank test. Note that adding observations that are censored before the first exact time has no effect. Also, because the test statistic is formally the same as the Mantel-Haenszel statistic, you can use proc FREQ with the CMH option (e.g., macro LogRank in the table below) rather than proc LIFETEST to perform log rank tests. The benefit of doing so is that you can also perform 1 df tests for trend, using the SCORES=TABLE option and appropriate numeric values for the group variable.
and are appropriate for ordinal variables.

THE LOG RANK TEST AS ONE MEMBER OF A FAMILY OF RANK TESTS
The log rank test statistic can be generalized by assigning weights \( w_i > 0 \) to the distinct exact times rather than weighting them all equally:

\[
T = \sum_i w_i \left( D_i - L_0(D_i) \right).
\]

To approximate the covariance matrix of \( T \) under \( H_0 \), we use \( \sum_i w_i^2 \Sigma_0 \), and we approximate the distribution of

\[
T' \left( \sum_i w_i^2 \Sigma_0 \right)^{-1} T \text{ under } H_0 \text{ by a chi-squared distribution with } K - 1 \text{ df.}
\]

THE GENERALIZED WILCOXON TEST, THE TARONE-WARE FAMILY, AND THE HARRINGTON-FLEMING FAMILY
Weighting the individual exact time (observed − expected under \( H_0 \)) vectors by the number at risk just before each exact time, \( r_1, \ldots, r_K = r_i \), gives the version of the generalized Wilcoxon test implemented in proc LIFETEST. Using weights equal to \( w_i^\gamma \) for \( 0 \leq \gamma \leq 1 \) gives a family of tests with the log rank test at one end of the range and the generalized Wilcoxon test at the other; Tarone and Ware (1977) suggest using \( \gamma = 1/2 \)

\[
\left( w_i = \sqrt{r_i} \right)
\]

as a kind of compromise with good power in a wide range of situations. Another weighting system, proposed by Harrington and Fleming (1982), is based on a pooled estimate of the survival function at each exact time:

\[
w_i = \left[ \hat{S}(t_i) \right]^{\rho} \text{ for some } \rho \geq 0 \text{. A natural choice for } \hat{S} \text{ is the Kaplan-Meier estimator. Cantor (2003) describes a macro that can help you pre-select the test most sensitive to the pattern of group differences you anticipate.}

WILL THE REAL GENERALIZED WILCOXON TEST PLEASE STEP FORWARD?
The name “generalized Wilcoxon” could lead you to think that if you feed the same uncensored data to LIFETEST and NPAR1WAY (with the Wilcoxon option), you’ll get the same chi-squared statistic and p-value. This is not the case. The \( T \) vector will be the same, but its estimated covariance matrix under \( H_0 \) will be different. Macro Gehan (see table below) implements the generalization of the Wilcoxon test to right-censored data proposed by Gehan in 1965, by calculating the permutation variance under \( H_0 \) (equal survival distributions, equal censoring distributions). Gehan’s test is the real generalized Wilcoxon test, in the sense that it is identical to the Wilcoxon test when no times are censored. To complicate the picture further, other covariance matrix estimators have been proposed. You might speculate that the versions that assume equal censoring patterns have better small-sample properties when censoring patterns are in fact equal. However, Gehan’s chi-squared statistic is affected by adding observations censored before the first exact time, a somewhat disconcerting property.

STRATIFIED RANK TESTS
For any test \( T \) in the family and its covariance matrix or estimated covariance matrix under \( H_0 \), say \( W \), given data for \( S \) strata we can calculate

\[
\left( \sum_i T_i \right)' \left( \sum_i W_i \right)^{-1} \left( \sum_i T_i \right)
\]

and approximate its null distribution by a chi-squared distribution with \( K - 1 \) df (Kalbfleisch and Prentice 2002). Of course, although we may think of this as testing equality of survival curves across treatment groups within each of the strata, differences among the treatment groups that are not similar across the strata (i.e., that tend to cancel when summed across strata) are less likely to be detected.

WHAT TO DO IF YOU HAVE VERSION 8
Log rank tests only – proc PHREG with TIES=DISCRETE
Create indicator variables for groups 2 through \( K \) and use statements like these:

```latex
class g1 g2 g3 g4;
proc phreg data=work ;
model time*death(0) = g2 g3 g4 / ties=discrete ;
strata diseaseStage ;
run ;
```

The score test p-value in the output is the log rank p-value. For example, the log rank test chi-squared statistic and p-value for the breast cancer survival dataset given in Cantor (1997, Output 3.3, p. 78) are highlighted below:

```
Testing Global Null Hypothesis: BETA=0

Likelihood Ratio Chi-Square DF Pr > ChiSq
15.6006 3 0.0014
Score 18.5232 3 0.0003
Wald 17.4993 3 0.0006
```

Two groups, no ties – proc LIFETEST TEST statement
Create an indicator variable for one of the groups and use it in the TEST statement. For instance, if you are comparing placebo and active treatment group survival and stratifying by disease stage, you could use statements like the ones below:

```latex
proc lifetest data=work method=PL plots=(s) ;
strata diseaseStage ;
time time*death(0) ;
test active ;
run ;
```

However, you rarely know in advance whether or not some of the times will be tied, which limits the usefulness of this trick. There is a Stata FAQ which may persuade you to avoid it:

Three or more groups, or ties – proc LIFETEST+ODS+SAS/IML

All we need to do is 1) capture the vector $T_i$ and covariance matrix $W_i$ for each stratum, 2) sum the vectors and their covariance matrices, and 3) calculate the chi-squared statistic. ODS makes the “capture” step simple:

1) ods output &prefix.HomCov=&prefix.HomCov (drop=RowName) HomStats=HomStats;
   proc lifetest data=&data method=KM plots=(s) ;
   by &by ;
   time &time*censor(&cenList) ;
   strata &strata ;
   run ;
   ods output close ;

The &prefix value is Wil for the generalized Wilcoxon test or Log for the log rank test. The stratum variable is &by and the group variable is &strata.

Step 2, “sum the vectors and covariance matrices,” could be completed without using IML, but step 3 requires a matrix inversion, so we may as well use IML for both steps. Before summing, we need the coefficients to use for the trend test:

use &data ;
read all var {&strata} into GroupVals ;
GroupVals = unique(GroupVals)` ;
if type( groupVals )='N' then do ;
c = groupVals ;
end ; else if type( groupVals )='C' then do ;
c = (1:nrow(groupVals))` ;
mattrib c rowname=(groupVals) ;
end ;
print, 'group coefficients for the trend test', c ;

If the group variable is numeric (say, dose), the macro will use its values as the coefficients, and if the group variable is character it will use the vector $(1,..., K)^T$. Check the output to make sure the character values have been numbered in the desired order.

Step 2 uses an IML module:

2) start Stratum ;
   setin HomStats ; * get O-E deviations vector ;
   read all var {&statistic} into dev where (&by=ByVal) ;
   read all var {&strata} into &strata where (&by=ByVal) ;
   setin &prefix.HomCov ; * get est. O-E covariance matrix ;
   read all into cov where (&by=ByVal) ;
   cov = cov[2:nrow(cov)] ; * drop 1st column=stratum ;
   t = dev`*ginv(cov)*dev ;
   print, "$\text{statistic stat. for this stratum}$, t ;
   if i > 1 then do ;
     sumDev = sumDev+dev ;
     sumCov = sumCov+cov ;
   end ; else do ;
     sumDev = dev ;
     sumCov = cov ;
   end ;
   finish Stratum ;

Step 3 could be carried out by discarding, say, the first row of the summed deviations vector and the first row and column of the summed covariance matrix to obtain an invertible matrix, but since IML has a generalized inverse function, we can save ourselves the trouble:

3) t = sumDev`*ginv(sumCov)*sumDev ;
   df = nrow(sumDev)-1 ;
   p = 1 - probchi(t, df) ;
   print, "$\text{statistic trend stat. for all strata}$, t df p ;
   results=t || df || p ;

If the group variable is numeric (say, dose), the macro will use its values as the coefficients, and if the group variable is character it will use the vector $(1,..., K)^T$. Check the output to make sure the character values have been numbered in the desired order.

Example: In a study (TEMPEST) evaluating tecadenoson for conversion of PSVT (paroxysmal supraventricular tachycardia) to normal heart rhythm, five tecadenoson dose regimens were evaluated, including (in order of increasing dose) C, D, and E. If PSVT persisted one minute after administration of study drug, a second dose was to be given, and if PSVT persisted one minute after the second dose, other means to terminate the PSVT, such as overdrive pacing or cardioversion (a kind of shock treatment), were permitted. Kaplan-Meier estimates of time to drug-related conversion if patients had been given the first dose only are shown below, by type of PSVT. (PSVT is caused by a kind of electrical short circuit within the heart; AVRT and AVNRT identify the type of short circuit.)

![Graph 1](image1)

PSVTTYPE=AVNRT

Survival Distribution Function

0.00 0.25 0.50 0.75 1.00

Time to conversion (sec)

0 100 200 300 400 500 600 700

STRATA: TRTCD=C Censored TRTCD=D Censored TRTCD=E Censored

PSVTTYPE=AVRT

Survival Distribution Function

0.00 0.25 0.50 0.75 1.00

Time to conversion (sec)

0 100 200 300 400 500 600 700

STRATA: TRTCD=C Censored TRTCD=D Censored TRTCD=E

Note that in both strata, all three curves are similar early and begin to separate later. The generalized Wilcoxon test gives greater weight to the early times than the log rank test, so the log
rank test would be expected to be more sensitive to differences like these. Log rank p-values calculated by the SlogRank macro (see table below) are shown below:

<table>
<thead>
<tr>
<th>Statistic for all strata</th>
<th>T</th>
<th>DF</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogR 8.5240139</td>
<td>2</td>
<td></td>
<td>0.014094</td>
</tr>
<tr>
<td>LogR trend statistic for all strata</td>
<td>T</td>
<td>DF</td>
<td>P</td>
</tr>
<tr>
<td>5.9008556</td>
<td>1</td>
<td></td>
<td>0.0151335</td>
</tr>
</tbody>
</table>

The corresponding generalized Wilcoxon p-values are:

<table>
<thead>
<tr>
<th>Statistic for all strata</th>
<th>T</th>
<th>DF</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilcoxon 4.9028178</td>
<td>2</td>
<td></td>
<td>0.0861721</td>
</tr>
<tr>
<td>Wilcoxon trend statistic for all strata</td>
<td>T</td>
<td>DF</td>
<td>P</td>
</tr>
<tr>
<td>4.4655367</td>
<td>1</td>
<td></td>
<td>0.0345852</td>
</tr>
</tbody>
</table>

As expected, the log rank test p-value is smaller than the Wilcoxon p-value, but if we had prespecified the Wilcoxon trend test, we would have been able to conclude that a dose-response trend had been established.

SPSS 8.0 computes the same p-values (Breslow = generalized Wilcoxon):

<table>
<thead>
<tr>
<th>Statistic for Equality of Survival Distributions for TRTCD Adjusted for TYPE</th>
<th>Statistic</th>
<th>df</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Rank</td>
<td>8.52</td>
<td>2</td>
<td>.0141</td>
</tr>
<tr>
<td>Breslow</td>
<td>4.90</td>
<td>2</td>
<td>.0862</td>
</tr>
</tbody>
</table>

WHAT TO DO WHEN YOU HAVE VERSION 9

When Release 9.9 (x=1?) arrives, life will become simpler for SAS programmers and statisticians performing survival analyses: a SUGI 28 paper by Rodriguez, Stokes, and Tobias indicates there will be a “new GROUP= option in the STRATA statement for performing stratified tests, Tarone-Ware, Peto-Peto, Fleming-Harrington in addition to logrank and Wilcoxon tests.” There is no mention of trend tests, but perhaps they’ll arrive in Release 9.(x+1).

WHEN TO STRATIFY? WHEN IS THE CHI-SQUARED APPROXIMATION GOOD ENOUGH?

All p-values for the tests above are chi-squared approximations that may or may not be adequate in small samples. Unfortunately, there are only a handful of papers on the small-sample properties of censored-data rank tests, so if you have, say, a “small” sample and a chi-squared p-value of .053, it may be time to invest in a package such as StatXact (www.cytel.com). The online documentation indicates that exact stratified tests are available for $K = 2$. Exact unstratified log rank tests can be performed using the Peto test in proc MULTTEST.

SUMMARY

If you are using version 8, you can use a SAS/STAT procedure to perform every type of test in the table below except:

<table>
<thead>
<tr>
<th>Macro</th>
<th>What it does</th>
<th>What it uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gehan</td>
<td>Implements Gehan’s extension of Wilcoxon’s rank sum test to right-censored data. Two groups only, unstratified.</td>
<td>Base SAS</td>
</tr>
<tr>
<td>HFrho</td>
<td>Implements Harrington-Fleming tests for right-censored data, without using IML. Two groups only, unstratified.</td>
<td>Base SAS, proc LIFETEST+ODS</td>
</tr>
<tr>
<td>Macro</td>
<td>What it does</td>
<td>What it uses</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>LogRank</td>
<td>Unstratified log rank test, including trend test, using proc FREQ with the CMH (Cochran-Mantel-Haenszel) option instead of proc LIFETEST. Macro illustrates the fact that the log rank test statistic is formally identical to the Mantel-Haenszel statistic for testing the null hypothesis of no association between the row and the column variable in a set of independent $2 \times K$ tables.</td>
<td>Base SAS, proc FREQ</td>
</tr>
<tr>
<td>SLogRank</td>
<td>Stratified log rank tests and stratified generalized Wilcoxon tests, including trend tests.</td>
<td>proc LIFETEST+ODS+SAS/IML</td>
</tr>
</tbody>
</table>

**MACROS AVAILABLE FROM WWW.SAS.COM**

<table>
<thead>
<tr>
<th>Macro</th>
<th>What it does</th>
<th>What it uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinRank (Cantor 1997)</td>
<td>Implements unstratified and stratified log rank, generalized Wilcoxon, Tarone-Ware, and Harrington-Fleming tests, including trend tests.</td>
<td>Base SAS, SAS/IML</td>
</tr>
</tbody>
</table>

**ACKNOWLEDGEMENTS**

Ai-Yu Wu and Lisa Meng provided the programs and statistical specifications for the PSVT example.

**REFERENCES**


The second edition (2003) is now available.


**CONTACT INFORMATION**

Ann Olmsted  
CV Therapeutics  
3172 Porter Drive  
Palo Alto, CA 94304  
650 384 8796  
fax 650 858 0390  
ann.olmsted@cvt.com