Title

stata.com

sts test — Test equality of survivor functions

Description

`sts test` tests the equality of survivor functions across two or more groups. The log-rank, Cox, Wilcoxon–Breslow–Gehan, Tarone–Ware, Peto–Peto–Prentice, and Fleming–Harrington tests are provided, in both unstratified and stratified forms.

`sts test` also provides a test for trend.

`sts test` can be used with single- or multiple-record or single- or multiple-failure st data.

Quick start

Log-rank test for the equality of survivor functions across levels of `v1` using `stset` data

```
sts test v1
```

Stratified log-rank test for equality of survivor functions across `v1` with strata `svar`

```
sts test v1, strata(svar)
```

As above, and show tests for each stratum

```
sts test v1, strata(svar) detail
```

Log-rank test for equality, and test for a trend in survivor functions for `v1`

```
sts test v1, trend
```

Test equality of survivor functions using the Wilcoxon–Breslow–Gehan test

```
sts test v1, wilcoxon
```

Likelihood-ratio test for the equality of survivor functions based on the Cox model

```
sts test v1, cox
```

Stratified Cox test of equality of survivor functions with strata `svar`

```
sts test v1, cox strata(svar)
```

Test equality of survivor functions using the Tarone–Ware test

```
sts test v1, tware
```

As above, and test for a trend using the same weights as used in the Tarone–Ware test

```
sts test v1, tware trend
```

Menu

Statistics > Survival analysis > Summary statistics, tests, and tables > Test equality of survivor functions

Syntax

```
sts test varlist [if] [in] [, options]
```

options	Description
Main	
`logrank`	perform log-rank test of equality; the default
`cox`	perform Cox test of equality
`wilcoxon`	perform Wilcoxon–Breslow–Gehan test of equality
`tware`	perform Tarone–Ware test of equality
`peto`	perform Peto–Peto–Prentice test of equality
`fh(`*p q*`)`	perform generalized Fleming–Harrington test of equality
`trend`	test trend of the survivor function across three or more ordered groups
`strata(`*varlist*`)`	perform stratified test on *varlist*, displaying overall test results
`detail`	display individual test results; modifies `strata()`
Options	
`mat(`*mname*$_1$ *mname*$_2$`)`	store vector $\mathbf{u}$ in *mname*$_1$ and matrix $\mathbf{V}$ in *mname*$_2$
`noshow`	do not show st setting information
`notitle`	suppress title

You must `stset` your data before using `sts test`; see [ST] **stset**.
Note that `fweights`, `iweights`, and `pweights` may be specified using `stset`; see [ST] **stset**.

Options

Main

`logrank`, `cox`, `wilcoxon`, `tware`, `peto`, and `fh(`*p q*`)` specify the test of equality desired. `logrank` is the default, unless the data are `pweighted`, in which case `cox` is the default and is the only possibility.

`wilcoxon` specifies the Wilcoxon–Breslow–Gehan test; `tware`, the Tarone–Ware test; `peto`, the Peto–Peto–Prentice test; and `fh()`, the generalized Fleming–Harrington test. The Fleming–Harrington test requires two arguments, p and q. When $p = 0$ and $q = 0$, the Fleming–Harrington test reduces to the log-rank test; when $p = 1$ and $q = 0$, the test reduces to the Mann–Whitney–Wilcoxon test.

`trend` specifies that a test for trend of the survivor function across three or more ordered groups be performed.

`strata(`*varlist*`)` requests that a stratified test be performed.

`detail` modifies `strata()`; it requests that, in addition to the overall stratified test, the tests for the individual strata be reported. `detail` is not allowed with `cox`.

Options

`mat(`*mname*$_1$ *mname*$_2$`)` requests that the vector $\mathbf{u}$ be stored in *mname*$_1$ and that matrix $\mathbf{V}$ be stored in *mname*$_2$. The other tests are rank tests of the form $\mathbf{u}'\mathbf{V}^{-1}\mathbf{u}$. This option may not be used with `cox`.

noshow prevents sts test from showing the key st variables. This option is seldom used because most people type stset, show or stset, noshow to set whether they want to see these variables mentioned at the top of the output of every st command; see [ST] **stset**.

notitle requests that the title printed above the test be suppressed.

Remarks and examples

stata.com

Remarks are presented under the following headings:

sts test tests the equality of the survivor function across groups. With the exception of the Cox test, these tests are members of a family of statistical tests that are extensions to censored data of traditional nonparametric rank tests for comparing two or more distributions. A technical description of these tests can be found in the *Methods and formulas* section of this entry. Simply, at each distinct failure time in the data, the contribution to the test statistic is obtained as a weighted standardized sum of the difference between the observed and expected number of deaths in each of the k groups. The expected number of deaths is obtained under the null hypothesis of no differences between the survival experience of the k groups.

The weights or weight function used determines the test statistic. For example, when the weight is 1 at all failure times, the log-rank test is computed, and when the weight is the number of subjects at risk of failure at each distinct failure time, the Wilcoxon–Breslow–Gehan test is computed.

The following table summarizes the weights used for each statistical test.

Test	Weight at each distinct failure time (t_i)
Log-rank	1
Wilcoxon–Breslow–Gehan	n_i
Tarone–Ware	$\sqrt{n_i}$
Peto–Peto–Prentice	$\widetilde{S}(t_i)$
Fleming–Harrington	$\widehat{S}(t_{i-1})^p\{1-\widehat{S}(t_{i-1})\}^q$

where $\widehat{S}(t_i)$ is the estimated Kaplan–Meier survivor-function value for the combined sample at failure time t_i, $\widetilde{S}(t_i)$ is a modified estimate of the overall survivor function described in *Methods and formulas*, and n_i is the number of subjects in the risk pool at failure time t_i.

These tests are appropriate for testing the equality of survivor functions across two or more groups. Up to 800 groups are allowed.

The "Cox" test is related to the log-rank test but is performed as a likelihood-ratio test (or, alternatively, as a Wald test) on the results from a Cox proportional hazards regression. The log-rank test should be preferable to what we have labeled the Cox test, but with pweighted data the log-rank test is not appropriate. Whether you perform the log-rank or Cox test makes little substantive difference with most datasets.

`sts test, trend` can be used to test against the alternative hypothesis that the failure rate increases or decreases as the level of the k groups increases or decreases. This test is appropriate only when there is a natural ordering of the comparison groups, for example, when each group represents an increasing or decreasing level of a therapeutic agent.

`trend` is not valid when `cox` is specified.

The log-rank test

`sts test`, by default, performs the log-rank test, which is, to be clear, the exponential scores test (Savage 1956; Mantel and Haenszel 1959; Mantel 1963, 1966). This test is most appropriate when the hazard functions are thought to be proportional across the groups if they are not equal.

This test statistic is constructed by giving equal weights to the contribution of each failure time to the overall test statistic.

In *Testing equality of survivor functions* in [ST] **sts**, we demonstrated the use of this command with the heart transplant data, a multiple-record, single-failure st dataset.

```
. use http://www.stata-press.com/data/r15/stan3
(Heart transplant data)
. sts test posttran
         failure _d:  died
   analysis time _t:  t1
                 id:  id

Log-rank test for equality of survivor functions

           |   Events         Events
posttran   |  observed       expected
-----------+-------------------------
0          |        30          31.20
1          |        45          43.80
-----------+-------------------------
Total      |        75          75.00
                chi2(1) =        0.13
                Pr>chi2 =      0.7225
```

We cannot reject the hypothesis that the survivor functions are the same.

`sts test, logrank` can also perform the stratified log-rank test. Say that it is suggested that calendar year of acceptance also affects survival and that there are three important periods: 1967–1969, 1970–1972, and 1973–1974. Therefore, a stratified test should be performed:

```
. stset, noshow
. generate group = 1 if year <= 69
(117 missing values generated)
. replace group=2 if year>=70 & year<=72
(78 real changes made)
. replace group=3 if year>=73
(39 real changes made)
```

```
. sts test posttran, strata(group)
Stratified log-rank test for equality of survivor functions
           |   Events         Events
posttran   |  observed    expected(*)
-----------+-------------------------
0          |         30          31.51
1          |         45          43.49
-----------+-------------------------
Total      |         75          75.00
(*) sum over calculations within group
                chi2(1) =       0.20
                Pr>chi2 =     0.6547
```

Still finding nothing, we ask Stata to show the within-stratum tests:

```
. sts test posttran, strata(group) detail
Stratified log-rank test for equality of survivor functions
-> group = 1
           |   Events         Events
posttran   |  observed       expected
-----------+-------------------------
0          |         14          13.59
1          |         17          17.41
-----------+-------------------------
Total      |         31          31.00
                chi2(1) =       0.03
                Pr>chi2 =     0.8558
-> group = 2
           |   Events         Events
posttran   |  observed       expected
-----------+-------------------------
0          |         13          13.63
1          |         20          19.37
-----------+-------------------------
Total      |         33          33.00
                chi2(1) =       0.09
                Pr>chi2 =     0.7663
-> group = 3
           |   Events         Events
posttran   |  observed       expected
-----------+-------------------------
0          |          3           4.29
1          |          8           6.71
-----------+-------------------------
Total      |         11          11.00
                chi2(1) =       0.91
                Pr>chi2 =     0.3410
-> Total
           |   Events         Events
posttran   |  observed    expected(*)
-----------+-------------------------
0          |         30          31.51
1          |         45          43.49
-----------+-------------------------
Total      |         75          75.00
(*) sum over calculations within group
                chi2(1) =       0.20
                Pr>chi2 =     0.6547
```

The Wilcoxon (Breslow–Gehan) test

`sts test, wilcoxon` performs the generalized Wilcoxon test of Breslow (1970) and Gehan (1965). This test is appropriate when hazard functions are thought to vary in ways other than proportionally and when censoring patterns are similar across groups.

The Wilcoxon test statistic is constructed by weighting the contribution of each failure time to the overall test statistic by the number of subjects at risk. Thus it gives heavier weights to earlier failure times when the number at risk is higher. As a result, this test is susceptible to differences in the censoring pattern of the groups.

`sts test, wilcoxon` works the same way as `sts test, logrank`:

```
. sts test posttran, wilcoxon

Wilcoxon (Breslow) test for equality of survivor functions
           |  Events       Events       Sum of
posttran   |  observed     expected      ranks
-----------+-------------------------------------
0          |        30        31.20        -85
1          |        45        43.80         85
-----------+-------------------------------------
Total      |        75        75.00          0

                chi2(1) =       0.14
                Pr>chi2 =     0.7083
```

With the `strata()` option, `sts test, wilcoxon` performs the stratified test:

```
. sts test posttran, wilcoxon strata(group)

Stratified Wilcoxon (Breslow) test for equality of survivor functions
           |  Events       Events       Sum of
posttran   |  observed     expected(*)  ranks(*)
-----------+-------------------------------------
0          |        30        31.51        -40
1          |        45        43.49         40
-----------+-------------------------------------
Total      |        75        75.00          0
(*) sum over calculations within group

                chi2(1) =       0.22
                Pr>chi2 =     0.6385
```

As with `sts test, logrank`, you can also specify the `detail` option to see the within-stratum tests.

The Tarone–Ware test

`sts test, tware` performs a test suggested by Tarone and Ware (1977), with weights equal to the square root of the number of subjects in the risk pool at time t_i.

Like Wilcoxon's test, this test is appropriate when hazard functions are thought to vary in ways other than proportionally and when censoring patterns are similar across groups. The test statistic is constructed by weighting the contribution of each failure time to the overall test statistic by the square root of the number of subjects at risk. Thus, like the Wilcoxon test, it gives heavier weights, although not as large, to earlier failure times. Although less susceptible to the failure and censoring pattern in the data than Wilcoxon's test, this could remain a problem if large differences in these patterns exist between groups.

`sts test, tware` works the same way as `sts test, logrank`:

```
. sts test posttran, tware

Tarone-Ware test for equality of survivor functions
           |  Events       Events       Sum of
posttran   |  observed     expected      ranks
-----------+------------------------------------
0          |        30        31.20  -9.3375685
1          |        45        43.80   9.3375685
-----------+------------------------------------
Total      |        75        75.00          0
                chi2(1) =       0.12
                Pr>chi2 =     0.7293
```

With the `strata()` option, `sts test, tware` performs the stratified test:

```
. sts test posttran, tware strata(group)

Stratified Tarone-Ware test for equality of survivor functions
           |  Events       Events       Sum of
posttran   |  observed   expected(*)    ranks(*)
-----------+------------------------------------
0          |        30        31.51  -7.4679345
1          |        45        43.49   7.4679345
-----------+------------------------------------
Total      |        75        75.00          0
(*) sum over calculations within group
                chi2(1) =       0.21
                Pr>chi2 =     0.6464
```

As with `sts test, logrank`, you can also specify the `detail` option to see the within-stratum tests.

The Peto–Peto–Prentice test

`sts test, peto` performs an alternative to the Wilcoxon test proposed by Peto and Peto (1972) and Prentice (1978). The test uses as the weight function an estimate of the overall survivor function, which is similar to that obtained using the Kaplan–Meier estimator. See *Methods and formulas* for details.

This test is appropriate when hazard functions are thought to vary in ways other than proportionally, but unlike the Wilcoxon–Breslow–Gehan test, it is not affected by differences in censoring patterns across groups.

`sts test, peto` works the same way as `sts test, logrank`:

```
. sts test posttran, peto

Peto-Peto test for equality of survivor functions
           |  Events       Events       Sum of
posttran   |  observed     expected      ranks
-----------+------------------------------------
0          |        30        31.20  -.86708453
1          |        45        43.80   .86708453
-----------+------------------------------------
Total      |        75        75.00          0
                chi2(1) =       0.15
                Pr>chi2 =     0.6979
```

With the `strata()` option, `sts test, peto` performs the stratified test:

```
. sts test posttran, peto strata(group)

Stratified Peto-Peto test for equality of survivor functions

           |   Events         Events        Sum of
posttran   |  observed    expected(*)      ranks(*)
-----------+-------------------------------------
0          |        30          31.51   -1.4212233
1          |        45          43.49    1.4212233
-----------+-------------------------------------
Total      |        75          75.00            0
(*) sum over calculations within group

                chi2(1) =       0.43
                Pr>chi2 =     0.5129
```

As with the previous tests, you can also specify the `detail` option to see the within-stratum tests.

The generalized Fleming–Harrington tests

`sts test, fh(`p q`)` performs the Harrington and Fleming (1982) class of test statistics. The weight function at each distinct failure time, t, is the product of the Kaplan–Meier survivor estimate at time $t-1$ raised to the p power and $1-$ the Kaplan–Meier survivor estimate at time $t-1$ raised to the q power. Thus, when specifying the Fleming and Harrington option, you must specify two nonnegative arguments, p and q.

When $p > q$, the test gives more weights to earlier failures than to later ones. When $p < q$, the opposite is true, and more weight is given to later failures than to earlier ones. When p and q are both zero, the weight is 1 at all failure times and the test reduces to the log-rank test.

`sts test, fh(`p q`)` works the same way as `sts test, logrank`. If we specify $p = 0$ and $q = 0$ we will get the same results as the log-rank test:

```
. sts test posttran, fh(0 0)

Fleming-Harrington test for equality of survivor functions

           |   Events         Events        Sum of
posttran   |  observed       expected        ranks
-----------+-------------------------------------
0          |        30          31.20   -1.1995511
1          |        45          43.80    1.1995511
-----------+-------------------------------------
Total      |        75          75.00            0

                chi2(1) =       0.13
                Pr>chi2 =     0.7225
```

We could, for example, give more weight to later failures than to earlier ones.

```
. sts test posttran, fh(0 3)

Fleming-Harrington test for equality of survivor functions

           |   Events         Events        Sum of
posttran   |  observed       expected        ranks
-----------+-------------------------------------
0          |        30          31.20   -.10288411
1          |        45          43.80    .10288411
-----------+-------------------------------------
Total      |        75          75.00            0

                chi2(1) =       0.01
                Pr>chi2 =     0.9065
```

Similarly to the previous tests, with the `strata()` option, `sts test, fh()` performs the stratified test:

```
. sts test posttran, fh(0 3) strata(group)

Stratified Fleming-Harrington test for equality of survivor functions

             |   Events        Events        Sum of
posttran     |  observed    expected(*)      ranks(*)
-------------+-----------------------------------------
0            |        30          31.51   -.05315105
1            |        45          43.49    .05315105
-------------+-----------------------------------------
Total        |        75          75.00             0
(*) sum over calculations within group

                chi2(1) =       0.00
                Pr>chi2 =     0.9494
```

As with the other tests, you can also specify the `detail` option to see the within-stratum tests.

The "Cox" test

The term *Cox test* is our own, and this test is a variation on the log-rank test using Cox regression.

One way of thinking about the log-rank test is as a Cox proportional hazards model on indicator variables for each of the groups. The log-rank test is a test that the coefficients are zero or, if you prefer, that the hazard ratios are one. The log-rank test is, in fact, a score test of that hypothesis performed on a slightly different (partial) likelihood function that handles ties more accurately.

Many researchers think that a (less precise) score test on the precise likelihood function is preferable to a (more precise) likelihood-ratio test on the approximate likelihood function used in Cox regression estimation. In our experience, it makes little difference:

```
. sts test posttran, cox

Cox regression-based test for equality of survival curves

             |   Events        Events       Relative
posttran     |  observed      expected         hazard
-------------+-----------------------------------------
0            |        30          31.20         0.9401
1            |        45          43.80         1.0450
-------------+-----------------------------------------
Total        |        75          75.00         1.0000

             LR chi2(1) =       0.13
                Pr>chi2 =     0.7222
```

By comparison, `sts test, logrank` also reported $\chi^2 = 0.13$, although the significance level was 0.7225, meaning that the χ^2 values differed in the fourth digit. As mentioned by Kalbfleisch and Prentice (2002, 20), a primary advantage of the log-rank test is the ease with which it can be explained to nonstatisticians, because the test statistic is the difference between the observed and expected number of failures within groups.

Our purpose in offering `sts test, cox` is not to promote its use instead of the log-rank test but to provide a test for researchers with sample-weighted data.

If you have sample weights (if you specified `pweights` when you `stset` the data), you cannot run the log-rank or Wilcoxon tests. The Cox regression model, however, has been generalized to sample-weighted data, and Stata's `stcox` can fit models with such data. In sample-weighted data, the likelihood-ratio statistic is no longer appropriate, but the Wald test based on the robust estimator of variance is.

Thus if we treated these data as sample-weighted data, we would obtain

```
. generate one = 1
. stset t1 [pw=one], id(id) time0(_t0) failure(died) noshow
                id:  id
     failure event:  died != 0 & died < .
obs. time interval:  (_t0, t1]
 exit on or before:  failure
            weight:  [pweight=one]
--------------------------------------------------------------------------
        172  total observations
          0  exclusions
--------------------------------------------------------------------------
        172  observations remaining, representing
        103  subjects
         75  failures in single-failure-per-subject data
  31,938.1  total analysis time at risk and under observation
                                        at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =     1,799
. sts test posttran, cox
Cox regression-based test for equality of survival curves
           |    Events         Events       Relative
posttran   |  observed       expected        hazard
-----------+-------------------------------------------
0          |     30.00          31.20        0.9401
1          |     45.00          43.80        1.0450
-----------+-------------------------------------------
Total      |     75.00          75.00        1.0000
          Wald chi2(1) =       0.13
               Pr>chi2 =     0.7181
```

`sts test, cox` now reports the Wald statistic, which is, to two digits, 0.13, just like all the others.

The trend test

When the groups to be compared have a natural order, such as increasing or decreasing age groups or drug dosage, you may want to test the null hypothesis that there is no difference in failure rate among the groups versus the alternative hypothesis that the failure rate increases or decreases as you move from one group to the next.

We illustrate this test with a dataset from a carcinogenesis experiment reprinted in Marubini and Valsecchi (1995, 126). Twenty-nine experimental animals were exposed to three levels (0, 1.5, 2.0) of a carcinogenic agent. The time in days to tumor formation was recorded. Here are a few of the observations:

```
. use http://www.stata-press.com/data/r15/marubini, clear
. list time event group dose in 1/9
```

	time	event	group	dose
1.	67	1	2	1.5
2.	150	1	2	1.5
3.	47	1	3	2
4.	75	0	1	0
5.	58	1	3	2
6.	136	1	2	1.5
7.	58	1	3	2
8.	150	1	2	1.5
9.	43	0	2	1.5

In these data, there are two variables that indicate exposure level. The `group` variable is coded 1, 2, and 3, indicating a one-unit separation between exposures. The `dose` variable records the actual exposure dosage. To test the null hypothesis of no difference among the survival experience of the three groups versus the alternative hypothesis that the survival experience of at least one of the groups is different, it does not matter if we use `group` or `dose`.

```
. stset time, fail(event) noshow
     failure event:  event != 0 & event < .
obs. time interval:  (0, time]
 exit on or before:  failure
--------------------------------------------------------------------------
         29  total observations
          0  exclusions
--------------------------------------------------------------------------
         29  observations remaining, representing
         15  failures in single-record/single-failure data
      2,564  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =       246
. sts test group
```

Log-rank test for equality of survivor functions

group	Events observed	Events expected
1	4	6.41
2	6	6.80
3	5	1.79
Total	15	15.00

```
           chi2(2) =       8.05
           Pr>chi2 =     0.0179
```

```
. sts test dose
```

Log-rank test for equality of survivor functions

dose	Events observed	Events expected
0	4	6.41
1.5	6	6.80
2	5	1.79
Total	15	15.00

```
          chi2(2) =       8.05
          Pr>chi2 =     0.0179
```

For the trend test, however, the distance between the values is important, so using `group` or `dose` will produce different results.

```
. sts test group, trend
```

Log-rank test for equality of survivor functions

group	Events observed	Events expected
1	4	6.41
2	6	6.80
3	5	1.79
Total	15	15.00

```
          chi2(2) =       8.05
          Pr>chi2 =     0.0179

Test for trend of survivor functions
          chi2(1) =       5.87
          Pr>chi2 =     0.0154
```

```
. sts test dose, trend
```

Log-rank test for equality of survivor functions

dose	Events observed	Events expected
0	4	6.41
1.5	6	6.80
2	5	1.79
Total	15	15.00

```
          chi2(2) =       8.05
          Pr>chi2 =     0.0179

Test for trend of survivor functions
          chi2(1) =       3.66
          Pr>chi2 =     0.0557
```

Although the above trend test was constructed using the log-rank test, any of the previously mentioned weight functions can be used. For example, a trend test on the data can be performed using the same weights as the Peto–Peto–Prentice test by specifying the `peto` option.

```
. sts test dose, trend peto
```

Peto-Peto test for equality of survivor functions

dose	Events observed	Events expected	Sum of ranks
0	4	6.41	-1.2792221
1.5	6	6.80	-1.3150418
2	5	1.79	2.5942639
Total	15	15.00	0

```
          chi2(2) =       8.39
          Pr>chi2 =     0.0150
Test for trend of survivor functions
          chi2(1) =       2.85
          Pr>chi2 =     0.0914
```

Video example

How to test the equality of survivor functions using nonparametric tests

Stored results

`sts test` stores the following in `r()`:

Scalars

`r(df)`	degrees of freedom	`r(chi2)`	χ^2
`r(df_tr)`	degrees of freedom, trend test	`r(chi2_tr)`	χ^2, trend test

Methods and formulas

Let $t_1 < t_2 < \cdots < t_k$ denote the ordered failure times; let d_j be the number of failures at t_j and n_j be the population at risk just before t_j; and let d_{ij} and n_{ij} denote the same things for group i, $i = 1, \ldots, r$.

We are interested in testing the null hypothesis

$$H_0 : \lambda_1(t) = \lambda_2(t) = \cdots = \lambda_r(t)$$

where $\lambda(t)$ is the hazard function at time t, against the alternative hypothesis that at least one of the $\lambda_i(t)$ is different for some t_j.

As described in Klein and Moeschberger (2003, 205–216), Kalbfleisch and Prentice (2002, 20–22), and Collett (2015, 50–51), if the null hypothesis is true, the expected number of failures in group i at time t_j is $e_{ij} = n_{ij}d_j/n_j$, and the test statistic

$$\mathbf{u}' = \sum_{j=1}^{k} W(t_j)(d_{1j} - e_{1j}, \ldots, d_{rj} - e_{rj})$$

is formed. $W(t_j)$ is a positive weight function defined as zero when n_{ij} is zero. The various test statistics are obtained by selecting different weight functions, $W(t_j)$. See the table in the *Remarks and examples* section of this entry for a list of these weight functions. For the Peto–Peto–Prentice test,

$$W(t_j) = \widetilde{S}(t_j) = \prod_{\ell : t_\ell \le t_j} \left(1 - \frac{d_\ell}{n_\ell + 1}\right)$$

The variance matrix $\mathbf{V}$ for $\mathbf{u}$ has elements

$$V_{il} = \sum_{j=1}^{k} \frac{W(t_j)^2 n_{ij} d_j (n_j - d_j)}{n_j (n_j - 1)} \left(\delta_{il} - \frac{n_{ij}}{n_j} \right)$$

where $\delta_{il} = 1$ if $i = l$ and $\delta_{il} = 0$, otherwise.

For the unstratified test, statistic $\mathbf{u}'\mathbf{V}^{-1}\mathbf{u}$ is distributed as χ^2 with $r - 1$ degrees of freedom.

For the stratified test, let $\mathbf{u}_s$ and $\mathbf{V}_s$ be the results of performing the above calculation separately within stratum, and define $\mathbf{u} = \sum_s \mathbf{u}_s$ and $\mathbf{V} = \sum_s \mathbf{V}_s$. The χ^2 test is given by $\mathbf{u}'\mathbf{V}^{-1}\mathbf{u}$ redefined in this way.

The "Cox" test is performed by fitting a (possibly stratified) Cox regression using `stcox` on $r - 1$ indicator variables, one for each group with one of the indicators omitted. The χ^2 test reported is then the likelihood-ratio test (no `pweights`) or the Wald test (based on the robust estimate of variance); see [ST] **stcox**.

The reported relative hazards are the exponentiated coefficients from the Cox regression renormalized, and the renormalization plays no role in calculating the test statistic. The renormalization is chosen so that the expected-number-of-failures-within-group weighted average of the regression coefficients is 0 (meaning that the hazard is 1). Let b_i, $i = 1, \ldots, r - 1$, be the estimated coefficients, and define $b_r = 0$. The constant K is then calculated with

$$K = \sum_{i=1}^{r} e_i b_i / d$$

where $e_i = \sum_j e_{ij}$ is the expected number of failures for group i, d is the total number of failures across all groups, and r is the number of groups. The reported relative hazards are $\exp(b_i - K)$.

The trend test assumes that there is natural ordering of the r groups, $r > 2$. Here we are interested in testing the null hypothesis

$$H_0 : \lambda_1(t) = \lambda_2(t) = \cdots = \lambda_r(t)$$

against the alternative hypothesis

$$H_a : \lambda_1(t) \leq \lambda_2(t) \leq \cdots \leq \lambda_r(t)$$

The test uses $\mathbf{u}$ as previously defined with any of the available weight functions. The test statistic is given by

$$\frac{\left(\sum_{i=1}^{r} a_i u_i \right)^2}{\mathbf{a}'\mathbf{V}\mathbf{a}}$$

where $a_1 \leq a_2 \leq \cdots \leq a_r$ are scores defining the relationship of interest. A score is assigned to each comparison group, equal to the value of the grouping variable for that group. $\mathbf{a}$ is the vector of these scores.

References

Breslow, N. E. 1970. A generalized Kruskal–Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika* 57: 579–594.

Collett, D. 2015. *Modelling Survival Data in Medical Research*. 3rd ed. Boca Raton, Fl: Chapman & Hall/CRC.

Gehan, E. A. 1965. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 52: 203–223.

Harrington, D. P., and T. R. Fleming. 1982. A class of rank test procedures for censored survival data. *Biometrika* 69: 553–566.

Kalbfleisch, J. D., and R. L. Prentice. 2002. *The Statistical Analysis of Failure Time Data*. 2nd ed. New York: Wiley.

Karrison, T. G. 2016. Versatile tests for comparing survival curves based on weighted log-rank statistics. *Stata Journal* 16: 678–690.

Klein, J. P., and M. L. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer.

Mantel, N. 1963. Chi-square tests with one degree of freedom; extensions of the Mantel–Haenszel procedure. *Journal of the American Statistical Association* 58: 690–700.

——. 1966. Evaluation of survival data and two new rank-order statistics arising in its consideration. *Cancer Chemotherapy Reports* 50: 163–170.

Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22: 719–748. Reprinted in *Evolution of Epidemiologic Ideas: Annotated Readings on Concepts and Methods*, ed. S. Greenland, pp. 112–141. Newton Lower Falls, MA: Epidemiology Resources.

Marubini, E., and M. G. Valsecchi. 1995. *Analysing Survival Data from Clinical Trials and Observational Studies*. New York: Wiley.

Peto, R., and J. Peto. 1972. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A* 135: 185–207.

Prentice, R. L. 1978. Linear rank tests with right censored data. *Biometrika* 65: 167–179.

Savage, I. R. 1956. Contributions to the theory of rank-order statistics—the two-sample case. *Annals of Mathematical Statistics* 27: 590–615.

Tarone, R. E., and J. H. Ware. 1977. On distribution-free tests for equality of survival distributions. *Biometrika* 64: 156–160.

White, I. R., S. Walker, and A. G. Babiker. 2002. strbee: Randomization-based efficacy estimator. *Stata Journal* 2: 140–150.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics* 1: 80–83.

Also see

[ST] **stcox** — Cox proportional hazards model

[ST] **sts** — Generate, graph, list, and test the survivor and cumulative hazard functions

[ST] **sts generate** — Create variables containing survivor and related functions

[ST] **sts graph** — Graph the survivor, hazard, or cumulative hazard function

[ST] **sts list** — List the survivor or cumulative hazard function

[ST] **stset** — Declare data to be survival-time data

[PSS] **power exponential** — Power analysis for the exponential test

[PSS] **power logrank** — Power analysis for the log-rank test