Significance Testing

Dave Harrington

May 14 - 18, 2018

Introduction

Comparing two distributions at a single time point

The log-rank test

Weighted log-rank tests

Tests for more than two groups

The stratified log-rank test

Derivations

Introduction

Significance testing with event-time data

In the medical literature, survival analysis is frequently used to analyze clinical trials that may potentially change practice.

- Significance testing is particularly important in this setting.

The figure on the next slide appeared at the beginning of Unit 1.

- Shows results of a randomized trial of ablation versus drug treatment for atrial fibrillation:
    - Estimates of probability of survival or hospital admission by treatment group
    - A *p*-value based on a *log-rank* test

This unit explores log-rank tests and other testing methods.

Example: Time to death or hospitalization

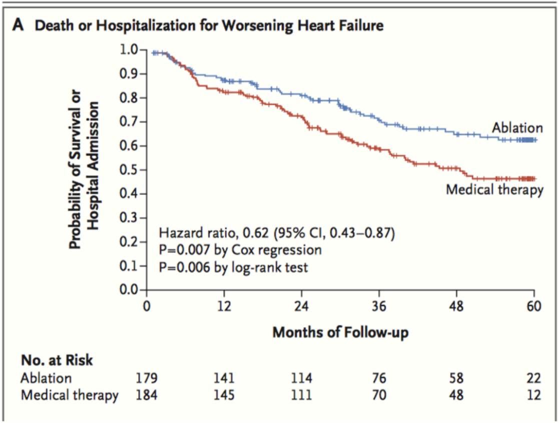


Figure 1: Figure from Marrouche, et al., *NEJM* 2018

Example: Clinical trial PBT01

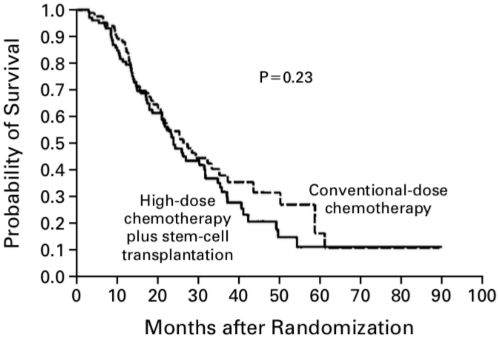


Figure 2: Figure from Stadtmauer, et al, NEJM 2000

The next slides reproduce this figure and *p*-value from patient-level data.

Numerical summary

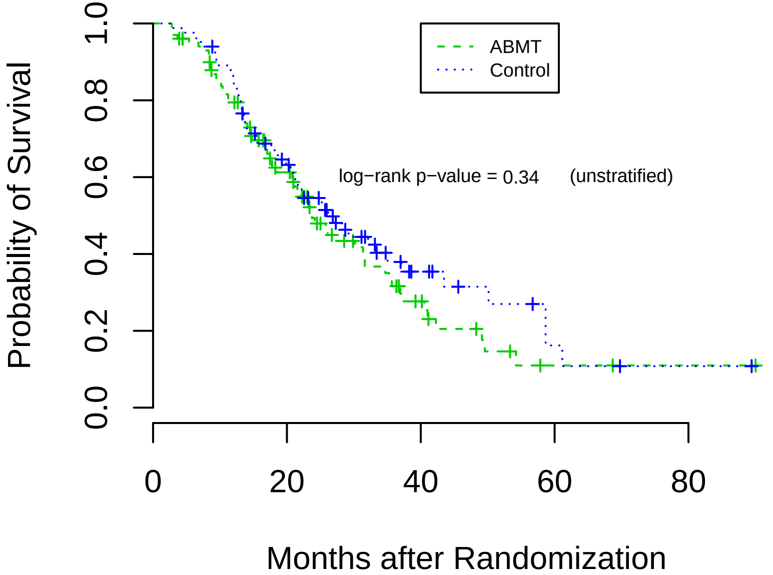
```r
library(survival)
library(eventtimedata)
data("pbt01")

pbt01.survival <- survfit(Surv(survival, died) ~ treatment,
                          data = pbt01)
pbt01.survival
```

Numerical summary . . .

```
## Call: survfit(formula = Surv(survival, died) ~ treatment, data = pbt01)
## 
##                      n events median 0.95LCL 0.95UCL
## treatment=abmt    101     64   23.7    20.8    31.6
## treatment=control  83     50   26.2    21.2    43.5
```

Figure

```r
library(survival)
library(eventtimedata)
data("pbt01")

pbt01.logrank.chisq = survdiff(Surv(survival, died) ~ treatment,
                                data = pbt01)$chisq
pbt01.logrank.pvalue = pchisq(pbt01.logrank.chisq, 1,
                              lower.tail = FALSE)

plot(pbt01.survival, lty = 2:3, col = 3:4, mark.time = TRUE,
     xlab = "Months after Randomization",
     ylab = "Probability of Survival",
     axes = FALSE,
     cex = 0.6)
legend(40, 1.0, c("ABMT", "Control"), lty = 2:3, col = 3:4,
       cex = 0.6)
text(40, 0.6, "log-rank p-value = ", cex = 0.6)
text(55, .6, round(pbt01.logrank.pvalue, digits = 2), cex = 0.6)
text(70, 0.6, "(unstratified)", cex = 0.6)
axis(1)
axis(2)
```



Note: The *p*-value is not identical to the one in the paper because the paper used a stratified test, stratifying on the cycle needed to induce complete response for the patient. Stratified tests coming later.

Parametric vs non-parametric approaches

In the medical literature, survival analysis is often used in the study of treatments for chronic diseases such as cancer, diabetes, or cardiovascular disease.

In most studies, a proportion of participants have not had an event by the time the study is analyzed.

- Thus, the right tail of the survival distribution is not observable.

Parametric approaches can be useful in some settings, but they assume a model for the entire curve, and extrapolate tail behavior.

Non-parametric methods make no assumptions about tail behavior and are less sensitive to outliers.

This section emphasizes *non-parametric methods*.

Two-sample non-parametric tests for comparing survival distributions

Comparing two distributions at a single time point

The log-rank test

Generalized Wilcoxon tests

The Fleming-Harrington family

References:

Hosmer & Lemeshow	Section 2.4
Collett	Section 2.5
Klein & Moeschberger	Section 7.3

Comparing two distributions at a single time point

Using $\widehat{S}_1(t)$ and $\widehat{S}_2(t)$

Sometimes a specific time point, $t^\star$, is of special interest.

- e.g., 5-year disease-free survival in cancer

Simple method:

- Use the independence and approximate normality of $\widehat{S}_k(t^\star); k \in \{0, 1\}$.
- Examine a confidence interval for the difference in estimated survival curves at $t^\star$.
- Reject $H_0 : S_1(t) = S_2(t)$ in favor of a two-sided alternative if the interval does not include 0.

Confidence interval for the difference of two survival curves

The 95% confidence interval is

$$\left[\left(\widehat{S}_1(t^\star) - \widehat{S}_0(t^\star)\right) \pm 1.96 \times \sqrt{V_1(t^\star) + V_0(t^\star)}\right],$$

where $V_k(t^\star)$ is the estimated variance of $\widehat{S}_k(t^\star)$.

This method is rarely used because

- it is not clear what $t^\star$ should be
- there is potential for abuse when applied *post-hoc*

Example

Use numerical estimates of the survival curves to find a 95% confidence interval for the difference in survival curves at time point $t^{\star} = 10$ weeks for the Cox and Oakes leukemia data.

The following slides show the estimates repeated from Unit 2 (Estimation).

KM numerical estimates, group == 0

```r
library(survival)
library(eventtimedata)
leukemia.group.0 =
  subset.data.frame(cox.oakes.leukemia, group == 0)
km.group.0 = survfit(Surv(time, relapse) ~ 1,
                     data = leukemia.group.0)
summary(km.group.0)
```

```
## Call: survfit(formula = Surv(time, relapse) ~ 1, data = leukemia.group.0)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     21       2   0.9048  0.0641      0.78754        1.000
##     2     19       2   0.8095  0.0857      0.65785        0.996
##     3     17       1   0.7619  0.0929      0.59988        0.968
##     4     16       2   0.6667  0.1029      0.49268        0.902
##     5     14       2   0.5714  0.1080      0.39455        0.828
##     8     12       4   0.3810  0.1060      0.22085        0.657
##    11      8       2   0.2857  0.0986      0.14529        0.562
##    12      6       2   0.1905  0.0857      0.07887        0.460
##    15      4       1   0.1429  0.0764      0.05011        0.407
##    17      3       1   0.0952  0.0641      0.02549        0.356
##    22      2       1   0.0476  0.0465      0.00703        0.322
##    23      1       1   0.0000     NaN           NA           NA
```

KM numerical estimates, group == 1

```
leukemia.group.1 =
  subset.data.frame(cox.oakes.leukemia, group == 1)
km.group.1 = survfit(Surv(time, relapse) ~ 1,
                     data = leukemia.group.1)
summary(km.group.1)
```

```
## Call: survfit(formula = Surv(time, relapse) ~ 1, data = leukemia.group.1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     6     21       3    0.857  0.0764        0.720        1.000
##     7     17       1    0.807  0.0869        0.653        0.996
##    10     15       1    0.753  0.0963        0.586        0.968
##    13     12       1    0.690  0.1068        0.510        0.935
##    16     11       1    0.627  0.1141        0.439        0.896
##    22      7       1    0.538  0.1282        0.337        0.858
##    23      6       1    0.448  0.1346        0.249        0.807
```

Calculations . . .

Lab exercise!

The log-rank test

Mantel-Haenszel log-rank test

The log-rank test is the most widely used non-parametric test.

Begin with a 2 × 2 table classifying those with and without the event of interest in a two group setting:

	Event		
Group	Yes	No	Total
0	d_0	$n_0 - d_0$	n_0
1	d_1	$n_1 - d_1$	n_1
Total	d	$n - d$	n

The table shows the observed numbers with and without events in each group, and the margin totals.

Mantel-Haenszel approach to a 2×2 table

Define D_0 as the random variable representing the number with an event in Group 0.

If the margins of this table $(d, n-d, n_0, n_1)$ are considered fixed, then D_0 follows a hypergeometric distribution, depending on one parameter (the population odds ratio, ψ).

Under H_0, the null hypothesis of no association between the event and group:

$$E(D_0) = \frac{n_0\, d}{n} = n_0 \left(\frac{d}{n}\right)$$

$$\mathrm{Var}(D_0) = \frac{n_0\, n_1\, d(n-d)}{n^2(n-1)}$$

Mantel-Haenszel approach to a 2×2 table...

Thus, the Mantel-Haenszel statistic is

$$\chi^2_{MH} = \frac{\left[d_0 - n_0\, d/n\right]^2}{\frac{n_0\, n_1\, d(n-d)}{n^2(n-1)}} \sim \chi^2_1$$

χ^2_{MH} is approximately equivalent to the Pearson χ^2 test for equality of the two groups given by:

$$\chi^2_p = \sum \frac{(o-e)^2}{e},$$

where o represents observed values and e the expected values.

Example: Toxicity in a clinical trial with two treatments

	Toxicity		
Group	Yes	No	Total
0	8	42	50
1	2	48	50
Total	10	90	100

$$\chi^2_p = 4.00 \quad (p = 0.046)$$

$$\chi^2_{MH} = 3.96 \quad (p = 0.047)$$

Pearson χ^2 vs MH

Note: the Pearson χ^2 test applies to the case where the row margins are fixed but not the column margins, as a test of equivalence between the proportions with events in the two groups.

In this case, the variance is slightly different:

$$\mathrm{Var}(D_0) = \frac{n_0\, n_1\, d(n-d)}{n^3}$$

For the case of K tables

Now suppose there are K 2×2 tables, all independent.

The goal is to test for a common group effect $H_0 : \psi_j = \psi = 1$ versus $H_A : \psi \neq 1$.

The *Cochran-Mantel-Haenszel test* for a common odds ratio not equal to 1 can be written as:

$$\chi^2_{CMH} = \frac{\left[\sum_{j=1}^{K}(d_{0j} - n_{0j} \times d_j/n_j)\right]^2}{\sum_{j=1}^{K} n_{1j} n_{0j} d_j (n_j - d_j)/[n_j^2(n_j - 1)]}$$

This statistic is distributed approximately as χ^2_1.

K TABLES . . .

The subscript j refers to the j-th table:

	Event		
Group	Yes	No	Total
0	d_{0j}	$n_{0j} - d_{0j}$	n_{0j}
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
Total	d_j	$n_j - d_j$	n_j

Log-rank Test: Applying CMH to survival data

For the two-group *log-rank* test:

- Construct a 2×2 table at each distinct failure time.
- Compare the failure rates between the two groups, conditional on the number at risk in the groups.
- Combine the results from each table using the Cochran-Mantel-Haenszel test.

Formal notation for the log-rank test

Let $t_1, \ldots, t_K$ represent the K ordered, distinct failure times.

The table at the j-th failure time, is

	Die/Fail		
Group	Yes	No	Total
0	d_{0j}	$r_{0j} - d_{0j}$	r_{0j}
1	d_{1j}	$r_{1j} - d_{1j}$	r_{1j}
Total	d_j	$r_j - d_j$	r_j

where

- d_{0j} and d_{1j} are the number of failures in group 0 and 1, respectively, at the j-th failure time
- r_{0j} and r_{1j} are the number at risk in groups 0 and 1, at the j-th failure time

The log-rank test statistic formula

$$\chi^2_{\text{log-rank}} = \frac{\left[\sum_{j=1}^{K}(d_{0j} - r_{0j} \times d_j/r_j)\right]^2}{\sum_{j=1}^{K} \frac{r_{1j} r_{0j} d_j (r_j - d_j)}{[r_j^2 (r_j - 1)]}}$$

If the tables are all independent, then this statistic will have an approximate χ^2 distribution with 1 df.

Notes about log-rank tests

The log-rank statistic depends on ranks of event times only, that is, on the order in which events and censorings occur.

If there are no tied failure times between the two groups, then $d_j = 1$ and the log-rank statistic simplifies to

$$\chi^2_{\text{log-rank}} = \frac{[\sum_{j=1}^{K} (d_{0j} - \frac{r_{0j}}{r_j})]^2}{\sum_{j=1}^{K} r_{1j} r_{0j} / r_j^2}$$

The numerator can be interpreted as $[\sum (o - e)]^2$, where

- o is the observed number of deaths in a group, and e is the expected number, given the risk sets.
- The expected number equals the number of failures times the proportion at risk in the group.
- It does not matter which group is used for the sum.

Notes about the log-rank...

The $(o - e)$ terms in the numerator can be written as

$$\frac{r_{0j} r_{1j}}{r_j} (\widehat{\lambda}_{1j} - \widehat{\lambda}_{0j})$$

Solution as a lab problem!

Assumptions behind log-rank test

Censoring is independent.

- This assumption is made in nearly all survival methods.

The contributions to the statistic made by the 2×2 tables can be treated as independent.

- Proven true in the 1980's

The log-rank test is most powerful when hazards have a constant ratio over time.

- This is termed the *proportional hazards* assumption.
- It is not required for validity under the null hypothesis.

Efficiency of the log-rank test

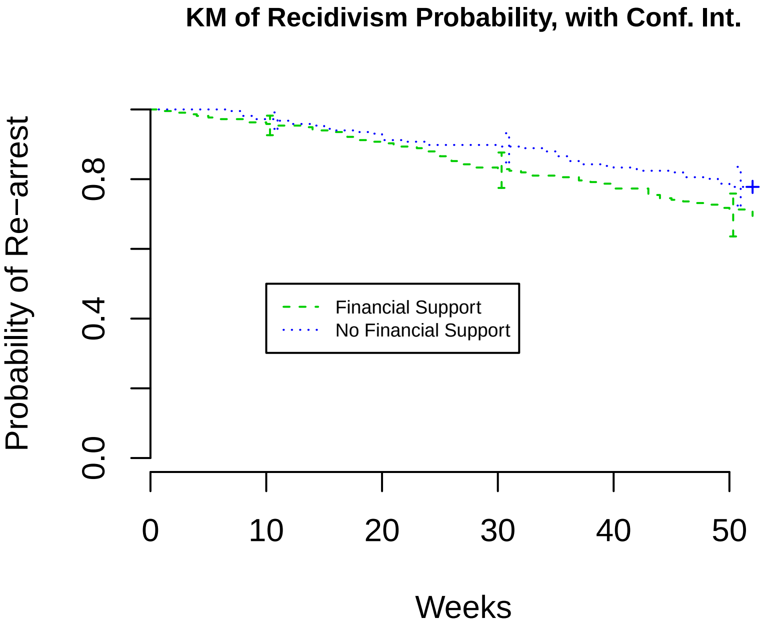
The CMH test for a series of tables stratified by a potential confounder is most powerful when . . .

- The tables have a constant odds ratio.

Analogously, the log-rank test is most powerful when . . .

- The hazard ratios are constant across t time intervals.
- This corresponds to *proportional hazards*.

The recidivism data



The recidivism data . . .

```r
library(survival)
library(eventtimedata)
data("rossi")
survfit(Surv(week, arrest) ~ fin,
                        data = rossi)
```

```
## Call: survfit(formula = Surv(week, arrest) ~ fin, data = rossi)
##
##           n events median 0.95LCL 0.95UCL
## fin=no  216     66     NA      NA      NA
## fin=yes 216     48     NA      NA      NA
```

```r
survdiff(Surv(week, arrest) ~ fin,
                        data = rossi)
```

```
## Call:
## survdiff(formula = Surv(week, arrest) ~ fin, data = rossi)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## fin=no  216       66     55.6      1.96      3.84
## fin=yes 216       48     58.4      1.86      3.84
##
##  Chisq= 3.8  on 1 degrees of freedom, p= 0.0501
```

What does non-proportional hazards look like?

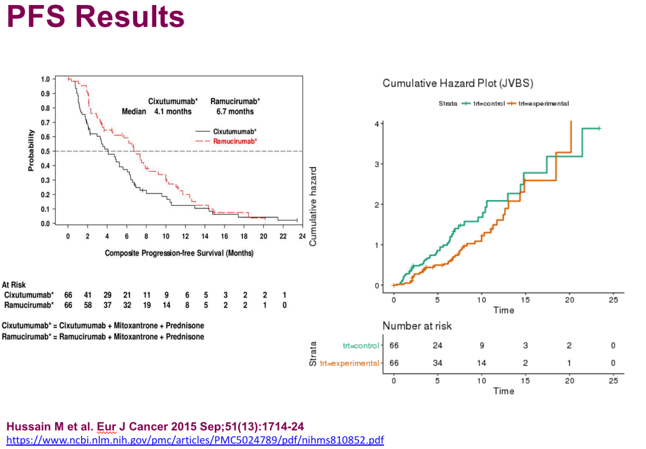
The next slides show figures presented at a February 5, 2018 meeting on non-proportional hazards.

All use data from published papers.

The workshop (sponsored by Duke University Margolis Center)

- reviewed instances where non-proportional hazards occurred in studies designed for drug approval
- discussed strategies for modifying usual methods of analysis

Randomized trial in prostate cancer



PFS Results

At Risk

Cixutumumab*	66	41	29	21	11	9	6	5	3	2	2	1
Ramucirumab*	66	58	37	32	19	14	8	5	2	2	1	0

Cixutumumab* = Cixutumumab + Mitoxantrone + Prednisone
Ramucirumab* = Ramucirumab + Mitoxantrone + Prednisone

Number at risk

Strata	0	5	10	15	20	25
trt=control	66	24	9	3	2	0
trt=experimental	66	34	14	2	1	0

Time

Hussain M et al. Eur J Cancer 2015 Sep;51(13):1714-24
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5024789/pdf/nihms810852.pdf

Figure 3: Data from Hussain, et al., *Euro J Cancer*, 2015

See Hussain, et al.

RCT in acute leukemia

INO-VATE trial results

- The primary analysis for CR was highly significant ($p < 0.001$)
- For OS

Kantarjian et al., NEJM 2016

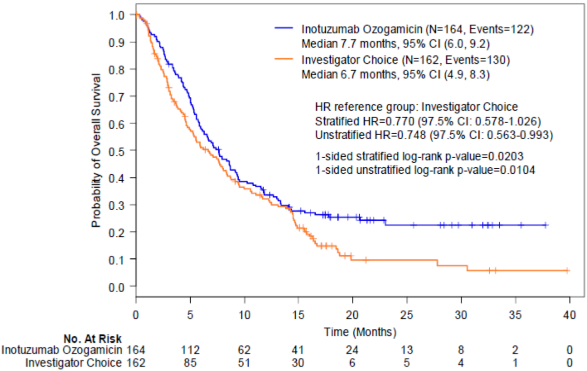


Figure 4: KM survival curves from Kantarjian, et al., *NEJM*, 2016

See Kantarjian, et al.

RCT in acute leukemia . . .

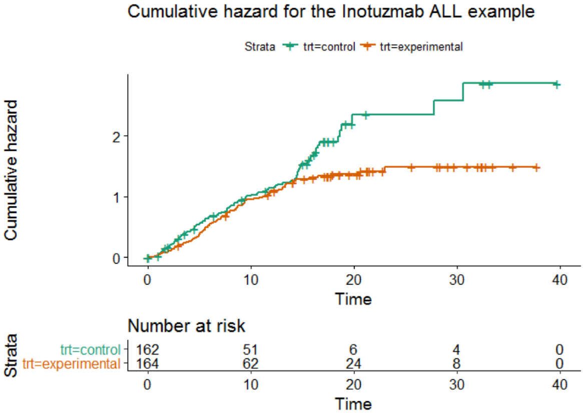


Figure 5: Cumulative hazards from Kantarjian, et al., *NEJM*, 2016

Another example, but with data

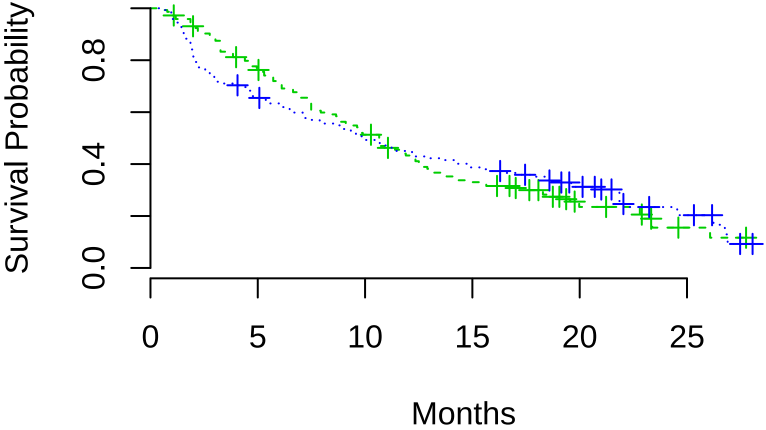
```r
#library(devtools)
#install_github("keaven/nphsim")
library(survival)
library(nphsim)
data(Ex6crossing)
survfit(Surv(month,evntd) ~ trt, data = Ex6crossing)
```

```
## Call: survfit(formula = Surv(month, evntd) ~ trt, data = Ex6crossing)
##
##         n events median 0.95LCL 0.95UCL
## trt=0 145    111  10.66    8.83    12.5
## trt=1 145    113   9.92    7.38    14.3
```

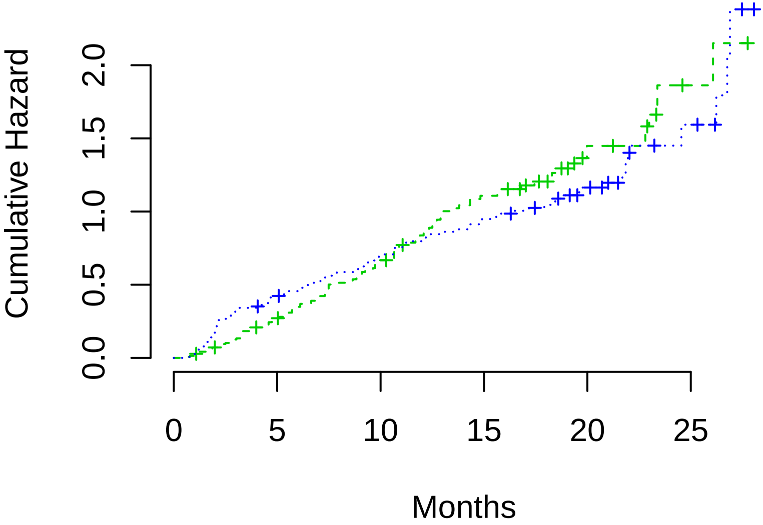
```r
survdiff(Surv(month,evntd) ~ trt, data = Ex6crossing)
```

```
## Call:
## survdiff(formula = Surv(month, evntd) ~ trt, data = Ex6crossing)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=0 145      111      110    0.0147    0.0296
## trt=1 145      113      114    0.0141    0.0296
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.863
```

The Kaplan-Meier estimate



The cumulative hazard



Weighted log-rank tests

The Tarone-Ware class of tests

This general class of tests is like the log-rank test, but adds weights w_j.

Many specific test statistics are included as special cases.

$$\chi^2_{TW} = \frac{[\sum_{j=1}^{K} w_j (d_{1j} - r_{1j} \times d_j / r_j)]^2}{\sum_{l=1}^{K} \frac{w_j^2 r_{1j} r_{0j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}}$$

Test statistic	Weight w_j
Log-rank	$w_j = 1$
Gehan's Wilcoxon	$w_j = r_j$
Peto/Prentice Wilcoxon	$w_j = n\widehat{S}(t_j)$
Fleming-Harrington	$w_j = [\widehat{S}(t_j)]^\rho \ [1 - \widehat{S}(t_j)]^\gamma$
Tarone-Ware	$w_j = \sqrt{r_j}$

r_j is the number of subjects at risk at the j^{th} event time.

Some background

The generalized Wilcoxon tests precede the Tarone-Ware or Fleming-Harrington class of tests.

- The Gehan-Wilcoxon was derived using a generalization of the U statistic approach to the Mann-Whitney-Wilcoxon.
- The Peto/Prentice Wilcoxon was derived using a generalization of linear rank statistics.

More details on the Fleming-Harrington test

The parameters ρ and γ can be any non-negative numbers:

- If $\rho = \gamma = 0$, $w_j = 1$ and the test is the usual log-rank test.
- If $\rho = 1$ and $\gamma = 0$, the test is similar to the Peto-Prentice.[1]
- If $\rho = 0$ and $\gamma = 1$, what happens to w_j over follow-up time?
- If $\rho = \gamma = 1$, the weight w_j reaches a maximum at the median, and is smaller for both large and small t_j.

The `survdiff()` function in R sets $\gamma = 0$ and allows the user to set ρ.

[1] This is the default "Fleming" test in SAS PROC LIFETEST.

Earlier numerical example, $\rho = 1$

This is a generalized Wilcoxon test

```
#library(devtools)
#install_github("keaven/nphsim")
library(survival)
library(nphsim)
data(Ex6crossing)
survdiff(Surv(month,evntd) ~ trt, rho = 1, data = Ex6crossing)
```

```
## Call:
## survdiff(formula = Surv(month, evntd) ~ trt, data = Ex6crossi
##     rho = 1)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=0 145     65.7     69.1     0.173     0.509
## trt=1 145     70.5     67.1     0.178     0.509
##
##  Chisq= 0.5  on 1 degrees of freedom, p= 0.476
```

Earlier numerical example, $\rho = 2$

```r
#library(devtools)
#install_github("keaven/nphsim")
library(survival)
library(nphsim)
data(Ex6crossing)
survdiff(Surv(month,evntd) ~ trt, rho = 2, data = Ex6crossing)
```

```
## Call:
## survdiff(formula = Surv(month, evntd) ~ trt, data = Ex6crossi
##     rho = 2)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=0 145     43.6     49.2     0.637      2.15
## trt=1 145     51.8     46.2     0.679      2.15
##
##  Chisq= 2.2  on 1 degrees of freedom, p= 0.142
```

Be careful with weighted LR tests

The weighted LR tests are often presented as emphasizing differences between two hazard functions.

For the F-H weights, $(\widehat{S}(t))^{\rho}(1 - \widehat{S}(t))^{\gamma}$,

- $\rho > 0, \ \gamma = 0$: weights early differences ($\rho = 1$ gen Wilcoxon)
- $\rho = 0, \ \gamma > 0$: weights late differences
- $\rho > 0, \ \gamma > 0$: weights differences near median
- $\rho = 0, \ \gamma = 0$: weights differences equally over time (log-rank)

Choosing the test post-hoc, based on observed data, leads to potentially increased Type I error.

The February 2018 Duke-Margolis workshop discussed ways to specify these tests in design and sample size calculations.

- Full workshop materials available at the link.

Tests for more than two groups

Introduction

Suppose data come from P different groups. The data from group p ($p = 1, \dots, P$) are:

$$(X_{p1}, \delta_{p1}) \dots (X_{pn_p}, \delta_{pn_p})$$

Tests are based on a $P \times 2$ table at each distinct K failure time.

- Compare the event rates between the P groups, conditional on the number at risk, combining the tables using the CMH approach
- Final test statistic has χ^2 distribution with $P - 1$ degrees of freedom

Example: prognosis in lymphoma

The data `lymphoma.prognosis` in the package `eventtimedata` was used as the training sample in the International Prognostic Index published by Shipp, et al. in 1993.

The data record survival time and censoring for 1,385 patients with non-Hodgkin's lymphoma treated at sites in the US, Canada, and Europe.

In this analysis, we look at the association of disease stage and survival. See the package documentation for the variable definitions.

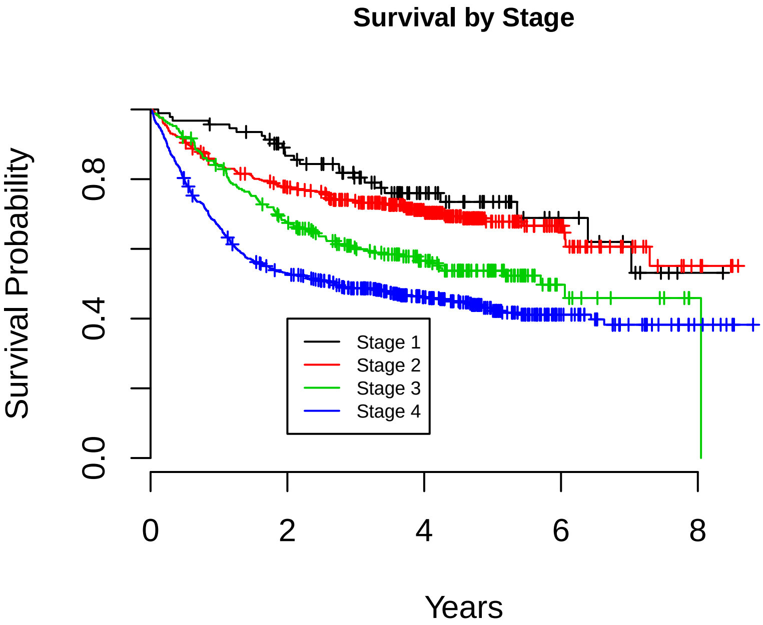
The code for the 4-group test

```r
library(survival)
library(eventtimedata)
data(lymphoma.prognosis)

stage.factor = as.factor(lymphoma.prognosis$STAGE)
died = lymphoma.prognosis$SURVIVAL - 1
died[died == 2] = 0  #recoding those lost to follow-up as censored
survival.time = lymphoma.prognosis$SURVTIME
lymphoma.survival <- survfit(Surv(survival.time, died) ~
                               stage.factor)
lymphoma.survival
survdiff(Surv(survival.time, died) ~ stage.factor)
```

The output for the 4-group test

```
## Call: survfit(formula = Surv(survival.time, died) ~ stage.factor)
##
##                  n events median 0.95LCL 0.95UCL
## stage.factor=1  93     24     NA    6.39      NA
## stage.factor=2 419    127     NA    7.30      NA
## stage.factor=3 253    112   5.70    4.11      NA
## stage.factor=4 620    340   2.68    1.84    4.22

## Call:
## survdiff(formula = Surv(survival.time, died) ~ stage.factor)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## stage.factor=1  93       24     48.6   12.4446    13.559
## stage.factor=2 419      127    201.0   27.2575    41.012
## stage.factor=3 253      112    114.4    0.0502     0.062
## stage.factor=4 620      340    239.0   42.6908    71.011
##
##  Chisq= 82.8  on 3 degrees of freedom, p= 0
```

The survival plot



The stratified log-rank test

Example: length of stay in a nursing home

The National Center for Health Services Research studied 36 for-profit nursing homes to assess

- effects of different financial incentives on length of stay

“Treated" nursing homes received

- Higher daily reimbursements for US Medicaid (financially needy) patients
- Bonuses for improving a patient’s health and sending them home

Study included 1601 patients admitted between May 1, 1981 and April 30, 1982.[2]

[2] Data are in `nursing.home` in the `eventtimedata` package.

Differences in length of stay by treatment

```
library(survival)
library(eventtimedata)
data(nursing.home)
survdiff(Surv(stay, cens) ~ rx, data = nursing.home)
```

```
## Call:
## survdiff(formula = Surv(stay, cens) ~ rx, data = nursing.home)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## rx=Control      889      684      677    0.0822     0.179
## rx=Intervention 712      595      602    0.0923     0.179
##
##  Chisq= 0.2  on 1 degrees of freedom, p= 0.672
```

A stratified analysis

Length of stay may also be associated with gender.

- Women tend to be healthier in the US.

A stratified test allows one to test for treatment differences, adjusting for gender (without using a modeling approach).

- assumes the shape of the hazard may vary between men and women, but that the effect of the incentive would be the same
- easy to do in almost any software

Differences in length of stay by treatment, stratified by gender

```r
library(survival)
library(eventtimedata)
data(nursing.home)
survdiff(Surv(stay, cens) ~ rx + strata(gender),
         data = nursing.home)
```

```
## Call:
## survdiff(formula = Surv(stay, cens) ~ rx + strata(gender), data = nu
##
##                    N Observed Expected (O-E)^2/E (O-E)^2/V
## rx=Control       889      684      679    0.0370    0.0812
## rx=Intervention 712      595      600    0.0418    0.0812
##
##  Chisq= 0.1  on 1 degrees of freedom, p= 0.776
```

The stratified test for the PBT01 data

Log-rank test stratified on `cycle.of.resp`.

This is the *p*-value in the Stadtmauer paper.

- Unstratified *p*-value (shown in earlier slides) is 0.34.

```r
library(survival)
library(eventtimedata)
data("pbt01")

survdiff(Surv(survival,died) ~ treatment + strata(cycle.of.resp),
              data = pbt01)
```

```
## Call:
## survdiff(formula = Surv(survival, died) ~ treatment + strata(cycle.of.resp),
##     data = pbt01)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## treatment=abmt    101       64     57.7     0.684      1.44
## treatment=control  83       50     56.3     0.702      1.44
##
##  Chisq= 1.4  on 1 degrees of freedom, p= 0.231
```

Derivations

The P-group log-rank statistic

Let $t_1, \ldots, t_K$ represent the K ordered, distinct failure times in the pooled sample.

At the j-th failure time, the following table summarizes the data,

	Fail		
Group	Yes	No	Total
1	d_{1j}	$r_{1j} - d_{1j}$	r_{1j}
.	.	.	.
P	d_{Pj}	$r_{Pj} - d_{Pj}$	r_{Pj}
Total	d_j	$r_j - d_j$	r_j

where d_{pj} is the number of deaths in group p at the j-th failure time, and r_{pj} is the number at risk at that time.

The tables are then combined using the CMH approach.

Details of the calculation

For one table at a particular failure time, the test statistic would be constructed from the $P \times 1$ vector of (observed - expected) values.

- Each group contributes one component of the sum.

Let $\mathbf{O}_j = (d_1, \ldots, d_{(P-1)j})^T$ be a vector of the observed number of failures in groups 1 to $(P-1)$ at the j-th death time. Given the risk sets $r_{1j}, \ldots, r_{Pj}$, and the fact that there are d_j deaths, $\mathbf{O}_j$ has mean

$$\mathbf{E}_j = \left(\frac{d_j r_{1j}}{r_j}, \ldots, \frac{d_j r_{(P-1)j}}{r_j} \right)^T$$

and variance-covariance matrix

$$\mathbf{V}_j = \begin{pmatrix} v_{11j} & v_{12j} & \ldots & v_{1(P-1)j} \\ & v_{22j} & \ldots & v_{2(P-1)j} \\ \ldots & & \ldots & \ldots \\ & & & v_{(P-1)(P-1)j} \end{pmatrix}$$

Details of the calculation. . .

- The ℓ-th diagonal element is:

$$v_{\ell\ell j} = r_{\ell j}(r_j - r_{\ell j})d_j(r_j - d_j)/[r_j^2(r_j - 1)]$$

- The ℓm-th off-diagonal element is:

$$v_{\ell m j} = r_{\ell j} r_{m j} d_j(r_j - d_j)/[r_j^2(r_j - 1)]$$

Details of the calculation . . .

The resulting χ^2 test for a single $P \times 1$ table has $(P-1)$ degrees of freedom and is constructed as follows:

$$(\mathbf{O}_j - \mathbf{E}_j)^T \, \mathbf{V}_j^{-1} \, (\mathbf{O}_j - \mathbf{E}_j)$$

To generalize to K tables (i.e., K failure times), combine as in the log-rank:

- Let $\mathbf{O}_j$, $\mathbf{E}_j$ and $\mathbf{V}_j$ with the sums over the K distinct failure times.
- That is, let $\mathbf{O} = \sum_{j=1}^{k} \mathbf{O}_j$, $\mathbf{E} = \sum_{j=1}^{k} \mathbf{E}_j$, and $\mathbf{V} = \sum_{j=1}^{k} \mathbf{V}_j$.

The test statistic is:

$$(\mathbf{O} - \mathbf{E})^T \, \mathbf{V}^{-1} \, (\mathbf{O} - \mathbf{E}),$$

and has a χ^2 distribution with $P-1$ degrees of freedom.

The stratified log-rank test

Used when assessing the association between survival and a factor X that has two different levels.

- Want to stratify by a second factor, that has S different levels.

First, divide the data into S separate groups.

Within group s ($s = 1, ..., S$),

- Construct the usual log-rank to assess the association between survival and the variable X.
- Let $t_{1s}, \ldots, t_{K_s s}$ represent the K_s ordered, distinct death times <u>in the s-th group</u>.

The stratified log-rank test . . .

At the j-th death time in group s:

	Die/Fail		
X	Yes	No	Total
1	d_{s1j}	$r_{s1j} - d_{s1j}$	r_{s1j}
2	d_{s2j}	$r_{s2j} - d_{s2j}$	r_{s2j}
Total	d_{sj}	$r_{sj} - d_{sj}$	r_{sj}

The stratified log-rank test . . .

Let

- O_s be the sum of the "o"s obtained by applying the log-rank calculations in the usual way to the data from group s.
- E_s be the sum of the "e"s,
- V_s be the sum of the "v"s.

The *stratified logrank* test statistic is

$$Z = \frac{\sum_{s=1}^{S}(O_s - E_s)}{\sqrt{\sum_{s=1}^{S}(V_s)}}$$

The test can easily be extended to weighted log-rank tests and to more than two levels of the factor X.