Chapter 3

Comparison of Survival Curves

3.1 Basis for Comparison of Survival Curves

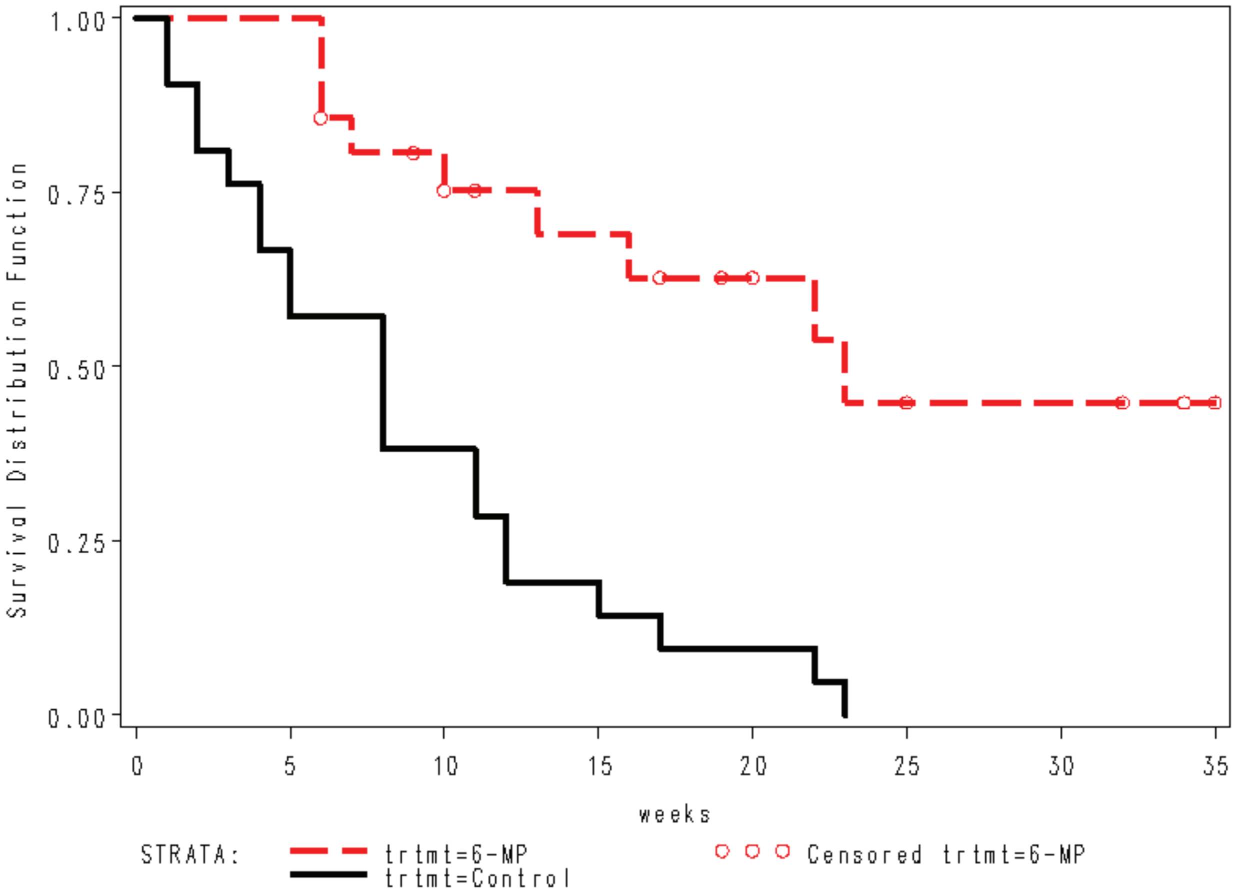
We have examined some nonparametric approaches for estimating the survival function, $\hat{S}(t)$, over time for a single sample of individuals.

Now we want to compare the survival estimates between two groups. How can we compare the two estimated survival distributions, $\hat{S}_1(t)$ and $\hat{S}_2(t)$?

For example: Time to relapse of leukemia patients

Figure 3.1: Kaplan-Meier curves for both trt groups

Leukemia study (Cox and Oakes)



Survival at a fixed time point

- **Sometimes a specific time point, t^*, is of special interest e.g. 5-year disease-free survival in cancer**
- **At this specific time point, is there a difference in the true survival? We could evaluate this question using our data by checking whether the "pointwise*" confidence intervals for the survival curves overlap at t^*?**
- **We can base a comparison on the approximate independent normal distributions of $\hat{S}_k(t^*); \quad k \in \{0,1\}$:**
- **In other words, we can check whether the 95% CI for the difference in survival estimates:**

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

includes 0, where $V_k(t^*)$ is the estimated variance of $\hat{S}_k(t^*)$.

* **Note: The pointwise confidence intervals we have been calculating correspond to a CI for $\hat{S}(t^*)$ at a particular point in time, t^*. The issue of confidence bands for the entire estimated survival function are discussed in Section 4.4 of Klein and Moeschberger**

Global comparisons of survival distributions

For testing

$$H_0 : S_1(t) = S_0(t) \text{ for every value of } t$$

Should we base the comparison on:

- the furthest distance between the two curves?
- the median survival for each group?
- the average hazard? (for exponential distributions, this would be like comparing the mean event times)
- adding up the difference between the two survival estimates over time?

$$\sum_j \left[\hat{S}(t_{jA}) - \hat{S}(t_{jB})\right]$$

- a weighted sum of differences, where the weights reflect the number at risk at each time?
- a rank-based test? i.e., we could rank all of the event times, and then see whether the sum of ranks for one group was less than the other.

Nonparametric comparisons of groups

All of the above are pretty reasonable options, and there have been several proposals for how to compare the survival of two groups. For the moment, we are sticking to nonparametric comparisons.

Why nonparametric?

- **fairly robust**
- **quite efficient relative to parametric tests**
- **often simple and intuitive**

Before continuing the description of the two-sample comparison, we give a more general perspective within which this approach is framed.

3.2 General Framework for Survival Analysis

We observe $(X_i, \delta_i, \mathbf{Z}_i)$ for individual i, where

- **X_i is a censored failure time random variable**
- **δ_i is the failure/censoring indicator**
- **$\mathbf{Z}_i$ represents a set of covariates**

Note that $\mathbf{Z}_i$ might be:

- **a scalar (a single covariate, say treatment or gender)**
- **or may be a $(p \times 1)$ vector (representing several different covariates).**

These covariates might be:

- continuous
- discrete
- time-varying (more later)

If $\mathbf{Z}_i$ is a scalar and is binary, then we are comparing the survival of two groups, like in the leukemia example.

More generally though, it is useful to build a model that characterizes the relationship between survival and all of the covariates of interest.

3.2.1 Relationships between covariates and survival outcomes

The general framework allows us to proceed in several different directions, as we start to evaluate the relationship between covariates (treatments or exposures) and survival outcomes:

- **Two group comparisons (e.g. logrank)**
- **Multigroup and stratified comparisons (e.g. stratified logrank)**
- **Failure time regression models**
  - **Cox proportional hazards model**
  - **Accelerated failure time model**

3.3 Two sample tests for Comparing Survival

- Mantel-Haenszel logrank test
- Peto & Peto's version of the logrank test
- Gehan's Generalized Wilcoxon
- Peto & Peto's and Prentice's generalized Wilcoxon
- Tarone-Ware and Fleming-Harrington classes
- Cox's F-test (non-parametric version)

References:

Hosmer & Lemeshow	Section 2.4
Collett	Section 2.5
Klein & Moeschberger	Section 7.3
Kleinbaum	Chapter 2
Lee	Chapter 5

3.3.1 Mantel-Haenszel Logrank test

The logrank test is the most well known and widely used.

It has an intuitive appeal, building on standard methods for binary data. (Later we will see that it can be obtained as the score test from a Cox Proportional Hazards model.)

First consider the following (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 previous table showed the observed numbers with and without events in each group, and the margin totals. But let's 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 distibution, depending on 1 parameter (the population odds ratio, ψ).

Under the null hypothesis of no association between the event and group, it follows that:

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

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

Therefore, under H_0: $$\chi^2_{MH} = \frac{[D_0 - n_0\, d/n]^2}{\frac{n_0\, n_1\, d(n-d)}{n^2(n-1)}} \sim \chi^2_1$$

This is the Mantel-Haenszel statistic and is approximately equivalent to the Pearson χ^2 test for equality of the two groups given by:

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

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

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

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

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 that case, the variance is slightly different than for the MH test:

$$\mathbf{Var}(d_0) = \frac{n_0\, n_1\, d(n-d)}{n^3}$$

Now suppose we have K (2×2) tables, all independent, and we want 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:

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

and this statistic is distributed approximately as χ^2_1. 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

How does this apply in survival analysis?

Suppose we observe

Group 1: $(X_{11}, \delta_{11}) \dots (X_{1n_1}, \delta_{1n_1})$

Group 0: $(X_{01}, \delta_{01}) \dots (X_{0n_0}, \delta_{0n_0})$

We could just count the numbers of failures in one of the groups:
eg., $D_0 = \sum_{j=1}^{K} \delta_{0j}$

Example: Leukemia data, just counting up the number of remissions in each treatment group.

	Fail		
Group	Yes	No	Total
0	21	0	21
1	9	12	21
Total	30	12	42

$$\chi^2_p = 16.8 \quad (p = 0.001) \qquad \chi^2_{MH} = 16.4 \quad (p = 0.001)$$

But, this does not account for the time at risk.

Conceptually, we would like to compare the KM survival curves. To do this, we first compare hazards at each failure time, and then aggregate over all the failure times.

Cox & Oakes Table 1.1 Leukemia example

Ordered	Group 0			Group 1		
Death Times	d_j	c_j	r_j	d_j	c_j	r_j
1	2	0	21	0	0	21
2	2	0	19	0	0	21
3	1	0	17	0	0	21
4	2	0	16	0	0	21
5	2	0	14	0	0	21
6	0	0	12	3	1	21
7	0	0	12	1	0	17
8	4	0	12	0	0	16
9	0	0	8	0	1	16
10	0	0	8	1	1	15
11	2	0	8	0	1	13
12	2	0	6	0	0	12
13	0	0	4	1	0	12
15	1	0	4	0	0	11
16	0	0	3	1	0	11
17	1	0	3	0	1	10
19	0	0	2	0	1	9
20	0	0	2	0	1	8
22	1	0	2	1	0	7
23	1	0	1	1	0	6
25	0	0	0	0	1	5

Logrank Test: Formal Definition

The logrank test can be obtained by constructing a (2×2) table at each distinct death time, and comparing the death rates between the two groups, conditional on the number at risk in the groups. The tables are then combined using the Cochran-Mantel-Haenszel test.

Let $t_1, ..., t_K$ represent the K ordered, distinct death times.
At the j-th death time, we have the following table:

	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 deaths in group 0 and 1, respectively at the j-th death time, and r_{0j} and r_{1j} are the number at risk at that time, in groups 0 and 1.

The logrank test is:

$$\chi^2_{logrank} = \frac{[\sum_{j=1}^{K}(D_{0j} - r_{0j} * d_j/r_j)]^2}{\sum_{j=1}^{K}\frac{r_{1j}r_{0j}d_j(r_j-d_j)}{[r_j^2(r_j-1)]}}$$

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

Based on the motivation for the logrank test, which of the survival-related quantities are we comparing at each time point?

- $\sum_{j=1}^{K} w_j \left[\hat{S}_1(t_j) - \hat{S}_2(t_j)\right]$ **?**
- $\sum_{j=1}^{K} w_j \left[\hat{\lambda}_1(t_j) - \hat{\lambda}_2(t_j)\right]$ **?**
- $\sum_{j=1}^{K} w_j \left[\hat{\Lambda}_1(t_j) - \hat{\Lambda}_2(t_j)\right]$ **?**

First several tables of leukemia data

```
CMH analysis of leukemia data

TABLE 1 OF TRTMT BY REMISS
CONTROLLING FOR FAILTIME=1

TRTMT     REMISS

Frequency|
Expected |       0|       1|  Total
---------+--------+--------+
       0 |      19 |       2 |      21
         |      20 |       1 |
---------+--------+--------+
       1 |      21 |       0 |      21
         |      20 |       1 |
---------+--------+--------+
Total          40        2       42



TABLE 2 OF TRTMT BY REMISS
CONTROLLING FOR FAILTIME=2

TRTMT     REMISS

Frequency|
Expected |       0|       1|  Total
---------+--------+--------+
       0 |      17 |       2 |      19
         |   18.05 |    0.95 |
---------+--------+--------+
       1 |      21 |       0 |      21
         |   19.95 |    1.05 |
---------+--------+--------+
Total          38        2       40
```

```
TABLE 3 OF TRTMT BY REMISS
CONTROLLING FOR FAILTIME=3

TRTMT     REMISS

Frequency|
Expected |       0|       1|  Total
---------+--------+--------+
       0 |      16 |       1 |      17
         | 16.553 | 0.4474 |
---------+--------+--------+
       1 |      21 |       0 |      21
         | 20.447 | 0.5526 |
---------+--------+--------+
Total          37        1       38



TABLE 4 OF TRTMT BY REMISS
CONTROLLING FOR FAILTIME=4

TRTMT     REMISS

Frequency|
Expected |       0|       1|  Total
---------+--------+--------+
       0 |      14 |       2 |      16
         | 15.135 | 0.8649 |
---------+--------+--------+
       1 |      21 |       0 |      21
         | 19.865 | 1.1351 |
---------+--------+--------+
Total          35        2       37
```

Ordered	Group 0		Combined				
Death Times	d_{0j}	r_{0j}	d_j	r_j	e_j	$o_j - e_j$	v_j
1	2	21	2	42	1.00	1.00	0.488
2	2	19	2	40	0.95	1.05	
3	1	17	1	38	0.45	0.55	
4	2	16	2	37	0.86	1.14	
5	2	14	2	35			
6	0	12	3	33			
7	0	12	1	29			
8	4	12	4	28			
10	0	8	1	23			
11	2	8	2	21			
12	2	6	2	18			
13	0	4	1	16			
15	1	4	1	15			
16	0	3	1	14			
17	1	3	1	13			
22	1	2	2	9			
23	1	1	2	7			
Sum						10.251	6.257

Calculating the logrank statistic by hand

Leukemia Example:

$o_j = d_{0j}$

$e_j = d_j r_{0j} / r_j$

$v_j = r_{1j} r_{0j} d_j (r_j - d_j) / [r_j^2 (r_j - 1)]$

$\sum_j (o_j - e_j) = 10.251$

$\sum_j v_j = 6.257$

$$\chi^2_{logrank} = \frac{\left(\sum_j (o_j - e_j)\right)^2}{\sum_j v_j} = \frac{(10.251)^2}{6.257} = 16.793$$

Notes about logrank tests:

- **The logrank statistic depends on ranks of event times only, eg, on the order in which events and censorings occur.**

- **If there are no tied deaths, then $d_j = 1$ and the logrank has the simplified form:**

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

- **Numerator can be interpreted as $\sum(o-e)$ where "o" is the observed number of deaths in group 0, and "e" is the expected number, given the risk sets. The expected number equals #deaths × proportion in group 0 at risk.**

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

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

- **It does not matter which group you choose to sum over.**

To see this, note that if we summed up (o-e) over the death times for the 6MP group we would get -10.251, and the sum of the variances is the same. So when we square the numerator, the test statistic is the same.

Power

Analogous to the CMH test for a series of tables at different levels of a confounder, the logrank test is most powerful when "odds ratios" are constant over time intervals. That is, it is most powerful for proportional hazards.

Checking the assumption of proportional hazards:

- **check to see if the estimated survival curves cross - if they do, then this is evidence that the hazards are <u>not</u> proportional**
- **more formal test: any ideas? We will come back to this later.**

What should be done if the hazards are NOT proportional?

- **If the difference between hazards has a consistent sign, the logrank test usually performs well.**
- **Other tests are available that are more powerful against different alternatives, e.g. weighted logrank tests, parametric tests, the Kolmogorov-Smirnoff non-parametric test.**

Getting the logrank statistic using SAS

- **We still use PROC LIFETEST**
- **Add a “STRATA” command, with treatment or exposure variable**
- **By default, the chi-square test is provided (2-sided)**
- **However, it also gives you the terms you need to calculate the 1-sided test; this is useful if we want to know which of the two groups has the higher estimated hazard over time.**
- **The STRATA command also gives the Gehan-Wilcoxon test (which we will talk about next)**

```
Title 'Cox and Oakes example';
data leukemia;
    input weeks remiss trtmt;
    cards;
6     0     1
6     1     1
6     1     1
6     1     1         /* data for 6MP group */
7     1     1
9     0     1
etc
1     1     0
1     1     0         /* data for placebo group */
2     1     0
2     1     0
etc
;

proc lifetest data=leukemia;
  time weeks*remiss(0);
  strata trtmt;
  title 'Logrank test for leukemia data';
run;
```

```
Logrank test for leukemia data

Summary of the Number of Censored and Uncensored Values

TRTMT       Total     Failed   Censored  %Censored

6-MP           21          9         12    57.1429
Control        21         21          0     0.0000

Total          42         30         12    28.5714

Testing Homogeneity of Survival Curves over Strata Time Variable
FAILTIME

Rank Statistics

TRTMT        Log-Rank     Wilcoxon

6-MP          -10.251      -271.00
Control        10.251       271.00

Covariance Matrix for the Log-Rank Statistics

TRTMT             6-MP       Control

6-MP           6.25696      -6.25696
Control       -6.25696       6.25696

Test of Equality over Strata
                               Pr >
Test      Chi-Square    DF   Chi-Square

Log-Rank     16.7929     1      0.0001  <== Here's the one we want!!
Wilcoxon     13.4579     1      0.0002
-2Log(LR)    16.4852     1      0.0001
```

Getting the Logrank test in Stata

Example using leukemia data (already saved as Stata dataset leukem.dat):

```
. use leukem
. stset remiss status
. sts test trt

         failure _d:  status
   analysis time _t:  remiss

Log-rank test for equality of survivor functions
------------------------------------------------
      |   Events          Events
trt   |  observed       expected
------+-------------------------
0     |        21          10.75
1     |         9          19.25
------+-------------------------
Total |        30          30.00

            chi2(1) =      16.79
            Pr>chi2 =     0.0000
```

[Note: interpret Pr>chi2=0.0000 to mean p<0.0001]

3.3.2 Linear Rank Tests

Linear rank tests are generalizations of the logrank test.

The logrank and other tests can be derived by assigning scores to the ranks of the death times, and are members of a general class of linear rank tests (for more detail, see Lee, ch 5)

First, define

$$\hat{\Lambda}(t) = \sum_{j:t_j \le t} \frac{d_j}{r_j}$$

where d_j and r_j are the number of deaths and the number at risk, respectively at the j-th ordered death time.

Then assign these scores (suggested by Peto and Peto):

Event	Score
Death at t_j	$w_j = 1 - \hat{\Lambda}(t_j)$
Censoring at t_j	$w_j = -\hat{\Lambda}(t_j)$

To calculate the logrank test, simply sum up the scores for group 0.

Example **Group 0: 15, 18, 19, 19, 20**

Group 1: 16+, 18+, 20+, 23, 24+

Calculation of logrank as a linear rank statistic

Ordered Data	Group	d_j	r_j	$\hat{\Lambda}(t_j)$	score w_j
15	0	1	10	0.100	0.900
16^+	1	0	9	0.100	-0.100
18	0	1	8	0.225	0.775
18^+	1	0	7	0.225	-0.225
19	0	2	6	0.558	0.442
20	0	1	4	0.808	0.192
20^+	1	0	3	0.808	-0.808
23	1	1	2	1.308	-0.308
24^+	1	0	1	1.308	-1.308

The logrank statistic S is sum of scores for group 0:

$$S = 0.900 + 0.775 + 0.442 + 0.442 + 0.192 = 2.75$$

Estimated variance of the logrank test

$$Var(S) = \frac{n_0 n_1 \sum_{j=1}^{n} w_j^2}{n(n-1)}$$

In this case, $Var(S) = 1.210$**, so**

$$Z = \frac{2.75}{\sqrt{1.210}} = 2.50 \implies \chi^2_{logrank} = (2.50)^2 = 6.25$$

Why is this form of the logrank equivalent?

The logrank statistic S is equivalent to $\sum(o - e)$ over the distinct death times, where "o" is the observed number of deaths in group 0, and "e" is the expected number, given the risk sets.

At deaths:	**weights are** $1 - \hat{\Lambda}$
At censorings:	**weights are** $-\hat{\Lambda}$

So we are summing up "1's" for deaths (to get d_{0j}), and subtracting $-\hat{\Lambda}$ at both deaths and censorings. This amounts to subtracting d_j/r_j at each death or censoring time in group 0, at or after the j-th death. Since there are a total of r_{0j} of these, we get $e = r_{0j} * d_j/r_j$.

Why is it called the logrank test?

Since $S(t) = \exp(-\Lambda(t))$, an alternative estimator of $S(t)$ is:

$$\hat{S}(t) = \exp(-\hat{\Lambda}(t)) = \exp(-\sum_{j:t_j<t} \frac{d_j}{r_j})$$

So, we can think of $\hat{\Lambda}(t) = -\log(\hat{S}(t))$ as yielding the "log-survival" scores used to calculate the statistic.

3.3.3 CMH-type Logrank versus the "Linear Rank" Logrank

A. CMH-type Logrank:

We motivated the logrank test through the CMH statistic for testing $H_o : OR = 1$ over K tables, where K is the number of distinct death times. This turned out to be what we get when we use the "STRATA" statement in SAS.

B. Linear Rank logrank:

The linear rank version of the logrank test is based on adding up "scores" for one of the two treatment groups. The particular scores that gave us the same logrank statistic were based on the Nelson-Aalen estimator, i.e., $\hat{\Lambda} = \sum \hat{\lambda}(t_j)$. This is what you get when you use the "TEST" statement in SAS.

Here are some comparisons, with a new example to show when the two types of logrank statistics will be equal.

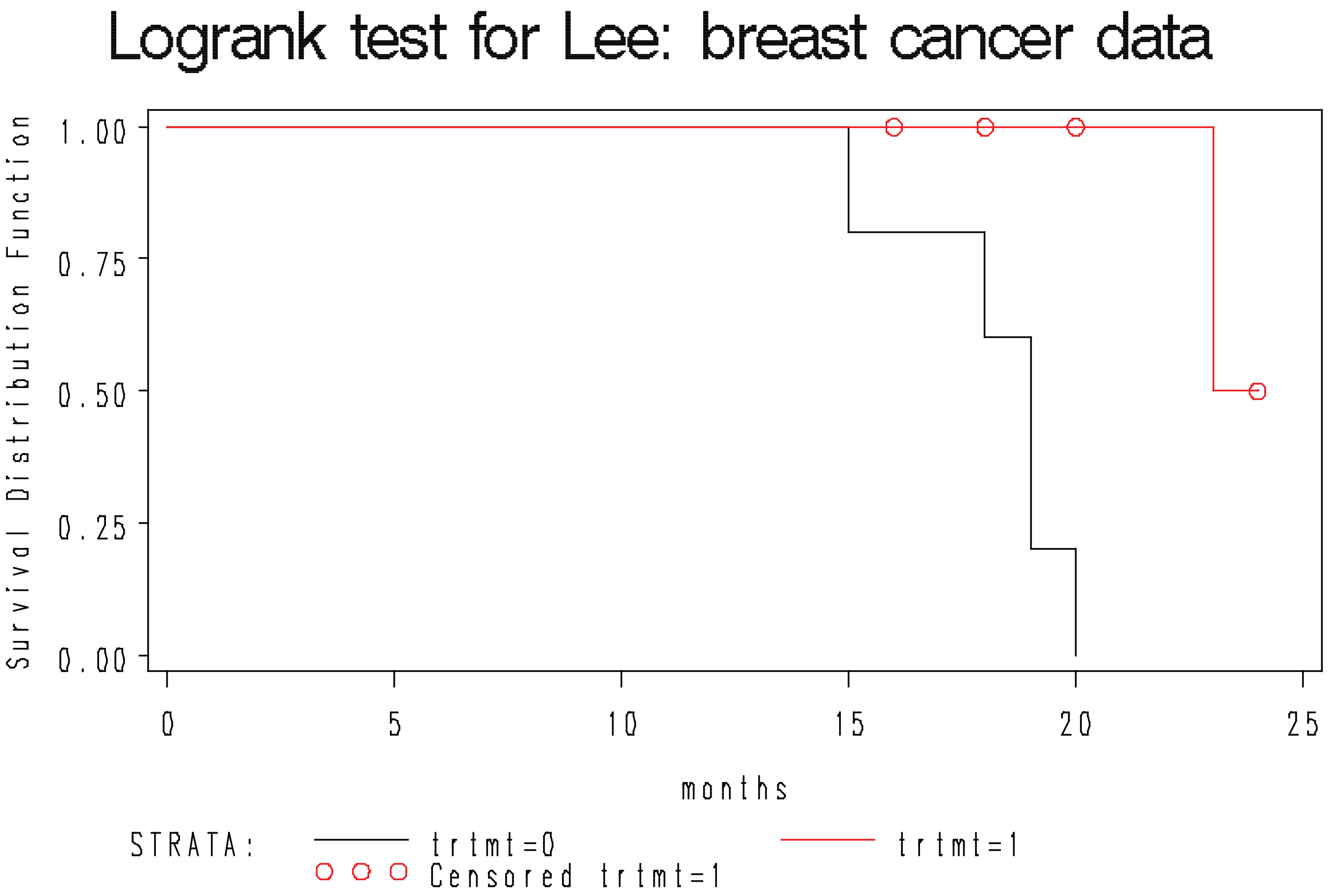
First, let's consider an example from Chapter 5 of Lee:

Ten female patients with breast cancer are randomized to receive either CMF (cyclic administration of cyclophosphamide, methatraxate, and fluorouracil) or no treatment after a radical mastectomy. At the end of two years these times to relapse have been recorded in months.

Example **Group 0: 15, 18, 19, 19, 20**

Group 1: 16+, 18+, 20+, 23, 24+

The Kaplan-Meier curves are shown next.


Logrank test for Lee: breast cancer data
Survival Distribution Function
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
months
STRATA:
trtmt=0
trtmt=1
Censored trtmt=1

A. The CMH-type logrank statistic:
(using the STRATA statement)

```
Rank Statistics

TRTMT        Log-Rank     Wilcoxon

Control        2.7500       18.000
Treated       -2.7500      -18.000

Covariance Matrix for the Log-Rank Statistics

TRTMT          Control       Treated

Control        1.08750      -1.08750
Treated       -1.08750       1.08750

Test of Equality over Strata
                               Pr >
Test      Chi-Square    DF  Chi-Square

Log-Rank      6.9540     1      0.0084
Wilcoxon      5.5479     1      0.0185
-2Log(LR)     3.3444     1      0.0674
```

This is exactly the same chi-square test that you would get if you calculated the numerator of the logrank as $\sum(o_j - e_j)$ and the individual variance terms at each failure time as $v_j = r_{1j}r_{0j}d_j(r_j - d_j)/[r_j^2(r_j - 1)]$

Ordered	Group 0		Combined				
Death Times	d_{0j}	r_{0j}	d_j	r_j	e_j	$o_j - e_j$	v_j
15	1	5	1	10	0.50	0.50	0.2500
18	1	4	1	8	0.50	0.50	0.2500
19	2	3	2	6	1.00	1.00	0.4000
20	1	1	2	4	0.25	0.75	0.1870
23	0	0	1	2	0.00	0.00	0.0000
Sum						2.75	1.0875

$$\chi^2_{logrank} = \frac{(2.75)^2}{1.0875} = 6.954$$

B. The “linear rank” logrank statistic: (using the TEST statement)

```
Univariate Chi-Squares for the LOG RANK Test

                 Test       Standard                   Pr >
Variable    Statistic      Deviation    Chi-Square    Chi-Square

GROUP          2.7500         1.0897        6.3684       0.0116


Covariance Matrix for the LOG RANK Statistics

Variable          TRTMT

TRTMT           1.18750
```

The test statistic is exactly the same for the linear rank and the CMH logrank tests, but the standard deviation is slightly different.

This is actually very close to what we would get if we use the Nelson-Aalen based "scores":

Calculation of logrank as a linear rank statistic

Ordered Data	Group	d_j	r_j	$\hat{\Lambda}(t_j)$	score w_j
15	0	1	10	0.100	0.900
16^+	1	0	9	0.100	-0.100
18	0	1	8	0.225	0.775
18^+	1	0	7	0.225	-0.225
19	0	2	6	0.558	0.442
20	0	1	4	0.808	0.192
20^+	1	0	3	0.808	-0.808
23	1	1	2	1.308	-0.308
24^+	1	1	1	1.308	-1.308
Sum(grp 0)					2.750

Note that the numerator is the exact same number (2.75) in both versions of the logrank test. The difference in the denominator is due to the way that ties are handled.

CMH-type variance:

$$\begin{aligned} var &= \sum \frac{r_{1j} r_{0j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)} \\ &= \sum \frac{r_{1j} r_{0j}}{r_j (r_j - 1)} \, \frac{d_j (r_j - d_j)}{r_j} \end{aligned}$$

Linear rank type variance:

$$var = \frac{n_0 n_1 \sum_{j=1}^{n} w_j^2}{n(n-1)}$$

An example where there are no tied death times

Example I **Group 0: 15, 18, 19, 21, 22**

Group 1: 16+, 17+, 20+, 23, 24+

A. The CMH-type logrank statistic: (using the STRATA statement)

```
Rank Statistics

TRTMT        Log-Rank     Wilcoxon

Control        2.5952       15.000
Treated       -2.5952      -15.000

Covariance Matrix for the Log-Rank Statistics

TRTMT           Control       Treated

Control         1.21712      -1.21712
Treated        -1.21712       1.21712


Test of Equality over Strata
                               Pr >
Test       Chi-Square    DF  Chi-Square

Log-Rank      5.5338      1      0.0187
Wilcoxon      4.3269      1      0.0375
-2Log(LR)     3.1202      1      0.0773
```

B. The "linear rank" logrank statistic: (using the TEST statement)

```
Univariate Chi-Squares for the LOG RANK Test

               Test      Standard                  Pr >
Variable    Statistic   Deviation   Chi-Square   Chi-Square

TRTMT          2.5952      1.1032       5.5338       0.0187


Covariance Matrix for the LOG RANK Statistics

Variable         TRTMT

TRTMT          1.21712
```

Note that this time, the variances of the two logrank statistics are exactly the same, equal to 1.217.

If there are no tied event times, then the two versions of the test will yield identical results. The more ties we have, the more it matters which version we use.

3.3.4 ‘Wilcoxon’ tests

Gehan’s Generalized Wilcoxon Test

First, let’s review the Wilcoxon test for uncensored data:
Denote observations from two samples by:

$$(X_1, X_2, \ldots, X_m) \quad \textbf{and} \quad (Y_1, Y_2, \ldots, Y_n)$$

Order the combined sample and define:

$$Z_{(1)} < Z_{(2)} < \cdots < Z_{(m+n)}$$

$$R_{i1} = \textbf{rank of } X_i$$

$$R_1 = \sum_{i=1}^{m+n} R_{i1}$$

Reject H_0 if R_1 is too big or too small, according to

$$\frac{R_1 - E(R_1)}{\sqrt{Var(R_1)}} \sim N(0,1)$$

where

$$\begin{aligned} E(R_1) &= \frac{m(m+n+1)}{2} \\ Var(R_1) &= \frac{mn(m+n+1)}{12} \end{aligned}$$

The Mann-Whitney form of the Wilcoxon is defined as:

$$U(X_i, Y_j) = U_{ij} = \begin{cases} +1 & \textbf{if} \quad X_i > Y_j \\ 0 & \textbf{if} \quad X_i = Y_j \\ -1 & \textbf{if} \quad X_i < Y_j \end{cases}$$

and

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}.$$

There is a simple correspondence between U and R_1:

$$\begin{aligned} R_1 &= m(m+n+1)/2 + U/2 \\ \textbf{so} \qquad U &= 2R_1 - m(m+n+1) \end{aligned}$$

Therefore,

$$\begin{aligned} E(U) &= 0 \\ Var(U) &= mn(m+n+1)/3 \end{aligned}$$

Extending Wilcoxon to censored data

The Mann-Whitney form leads to a generalization for censored data. Define

$$U(X_i, Y_j) = U_{ij} = \begin{cases} +1 & \textbf{if} \quad x_i > y_j \quad \textbf{or} \quad x_i^+ \geq y_j \\ 0 & \textbf{if} \quad x_i = y_i \quad \textbf{or lower value censored} \\ -1 & \textbf{if} \quad x_i < y_j \quad \textbf{or} \quad x_i \leq y_j^+ \end{cases}$$

Then define

$$W = \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}$$

Thus, there is a contribution to W for every comparison where both observations are failures (except for ties), or where a censored observation is greater than or equal to a failure.

Looking at all possible pairs of individuals between the two treatment groups makes this a nightmare to compute by hand!

Gehan found an easier way to compute the above. First, pool the sample of $(n+m)$ observations into a single group, then compare each individual with the remaining $n+m-1$: For comparing the i-th individual with the j-th, define

$$U_{ij} = \begin{cases} +1 & \textbf{if} \quad t_i > t_j \quad \textbf{or} \quad t_i^+ \geq t_j \\ -1 & \textbf{if} \quad t_i < t_j \quad \textbf{or} \quad t_i \leq t_j^+ \\ 0 & \qquad otherwise \end{cases}$$

Then

$$U_i = \sum_{j=1}^{m+n} U_{ij}$$

Thus, for the i-th individual, U_i is the number of observations which are definitely less than t_i minus the number of observations that are definitely greater than t_i. We assume censorings occur after deaths, so that if $t_i = 18^+$ and $t_j = 18$, then we add 1 to U_i.

The Gehan statistic is defined as

$$\begin{aligned} U &= \sum_{i=1}^{m+n} U_i \, \mathbf{1}_{\{i \textbf{ in group 0}\}} \\ &= W \end{aligned}$$

U **has mean 0 and variance**

$$var(U) = \frac{mn}{(m+n)(m+n-1)} \sum_{i=1}^{m+n} U_i^2$$

Example from Lee:

Group 0: **15, 18, 19, 19, 20**

Group 1: **16+, 18+, 20+, 23, 24+**

Time	Group	U_i	U_i^2
15	0	-9	81
16^+	1	1	1
18	0	-6	36
18^+	1	2	4
19	0	-2	4
19	0	-2	4
20	0	1	1
20^+	1	5	25
23	1	4	16
24^+	1	6	36
SUM		-18	208

Using these calculations we have:

$$
\begin{aligned}
U &= -18 \\
Var(U) &= \frac{(5)(5)(208)}{(10)(9)} \\
&= 57.78
\end{aligned}
$$

$$\textbf{and} \quad \chi^2 = (-18)^2/57.78 = 5.61$$

Obtaining the Gehan-Wilcoxon test in SAS

```
data leedata;
  infile 'lee.dat';
  input time cens group;

proc lifetest data=leedata;
  time time*cens(0);
  strata group; run;
```

SAS OUTPUT: Gehans Wilcoxon test

```
Rank Statistics
TRTMT        Log-Rank     Wilcoxon

Control        2.7500       18.000
Treated       -2.7500      -18.000

Covariance Matrix for the Wilcoxon Statistics

TRTMT          Control       Treated

Control        58.4000      -58.4000
Treated       -58.4000       58.4000

Test of Equality over Strata
                               Pr >
Test      Chi-Square    DF  Chi-Square

Log-Rank      6.9540     1      0.0084
Wilcoxon      5.5479     1      0.0185 **this is Gehan's test
-2Log(LR)     3.3444     1      0.0674
```

Notes about SAS Wilcoxon Test:

SAS calculates the Wilcoxon as $-U$ instead of U (sign = logrank sign). Also, SAS gets something slightly different for the variance, and this does not seem to depend on whether there are ties. E.g., the hypothetical dataset on p.6 without ties yields $U = -15$ and $\sum U_i^2 = 182$:

$$Var(U) = \frac{(5)(5)(182)}{(10)(9)} = 50.56 \quad \textbf{and} \quad \chi^2 = \frac{(-15)^2}{50.56} = 4.45$$

while SAS gives the following:

```
Rank Statistics

TRTMT        Log-Rank     Wilcoxon

Control        2.5952       15.000
Treated       -2.5952      -15.000

Covariance Matrix for the Wilcoxon Statistics

TRTMT          Control       Treated

Control        52.0000      -52.0000
Treated       -52.0000       52.0000

Test of Equality over Strata
                               Pr >
Test      Chi-Square    DF  Chi-Square

Log-Rank      5.5338     1      0.0187
Wilcoxon      4.3269     1      0.0375
-2Log(LR)     3.1202     1      0.0773
```

Why aren't they exactly the same?

There may be very slight differences in computational formulas that seem to make a difference for this example.

However, these examples only include 5 subjects per treatment arm.

In practice, we should not be applying a logrank test to such a small dataset! (we've used it here only to show computations)

Even though the test is non-parametric in terms of not making distributional assumptions, it still relies on a large enough sample size for the test statistics to be appropriate (the test statistics are "asymptotic" or "large sample" tests).

With larger sample sizes, the computational differences become negligible.

Obtaining the Gehan-Wilcoxon test in Stata

(same as before, but adding “Wilcoxon” option to test command)

Example: (leukemia data)

```
. stset remiss status
. sts test trt, wilcoxon

Wilcoxon (Breslow) test for equality of survivor functions
----------------------------------------------------------

      |  Events                                 Sum of
trt   |  observed        expected                ranks
------+-------------------------------------------------
0     |        21           10.75                  271
1     |         9           19.25                 -271
------+-------------------------------------------------
Total |        30           30.00                    0

            chi2(1) =       13.46
            Pr>chi2 =      0.0002
```

This is equivalent to the Gehan Wilcoxon test provided by SAS (p.27).

3.3.5 Generalized Wilcoxon: Peto & Peto, Prentice

For a death at t: **Score** $= \hat{S}(t+) + \hat{S}(t-) - 1$

For a censoring at t: **Score** $= \hat{S}(t+) - 1$

The test statistic is $\sum(scores)$ for group 0.

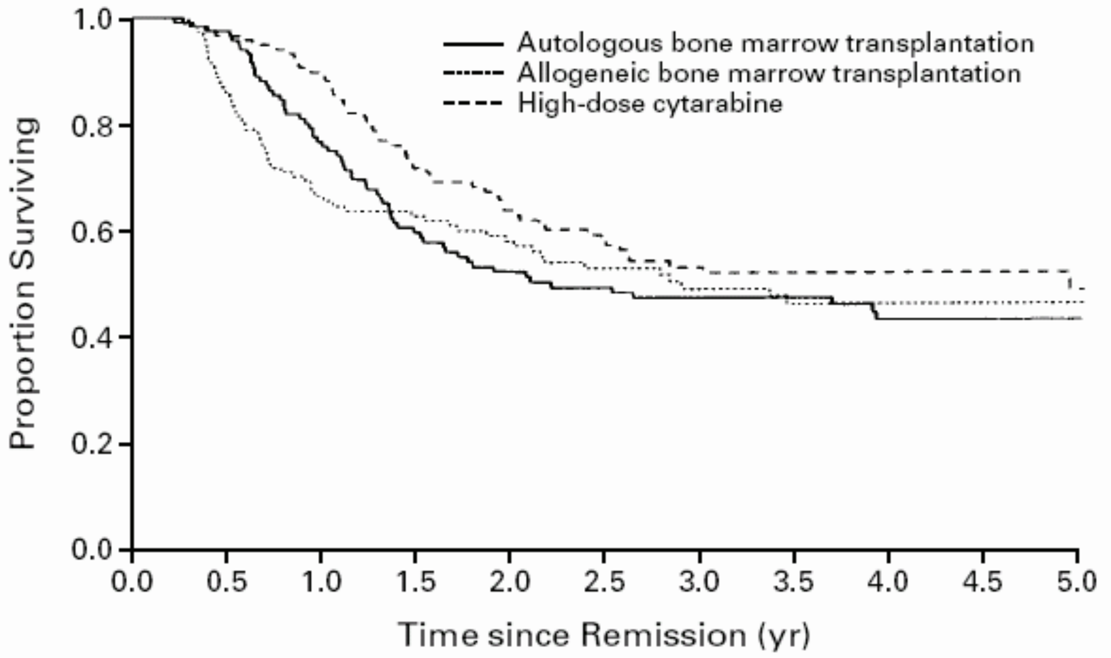
Time	Group	d_j	r_j	$\hat{S}(t+)$	score w_j
15	0	1	10	0.900	0.900
16^+	1	0	9	0.900	-0.100
18	0	1	8	0.788	0.688
18^+	1	0	7	0.788	-0.212
19	0	2	6	0.525	0.313
20	0	1	4	0.394	-0.081
20^+	1	0	3	0.394	-0.606
23	1	1	2	0.197	-0.409
24^+	1	0	1	0.197	-0.803

$$\sum w_j \mathbf{1}_{\{j \text{ in group } 0\}} = 0.900 + 0.688 + 2*(0.313) + (-0.081) = 2.13$$

$$Var(S) = \frac{n_0 n_1 \sum_{j=1}^{n} w_j^2}{n(n-1)} = 0.765$$

$$\text{so} \quad Z = 2.13/\sqrt{0.765} = 2.433$$

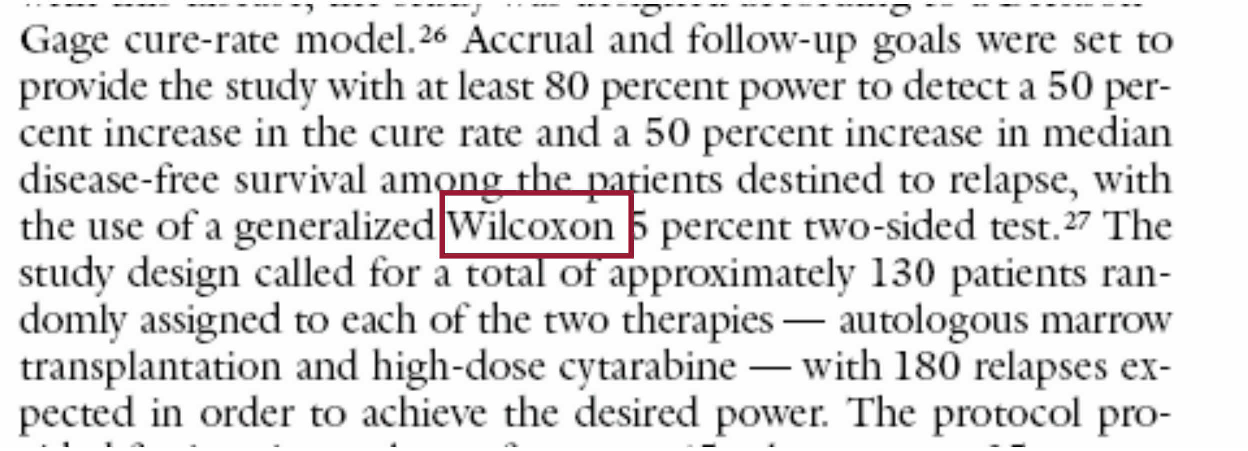
Chemo compared with bone marrow transplantation,NEJM'98



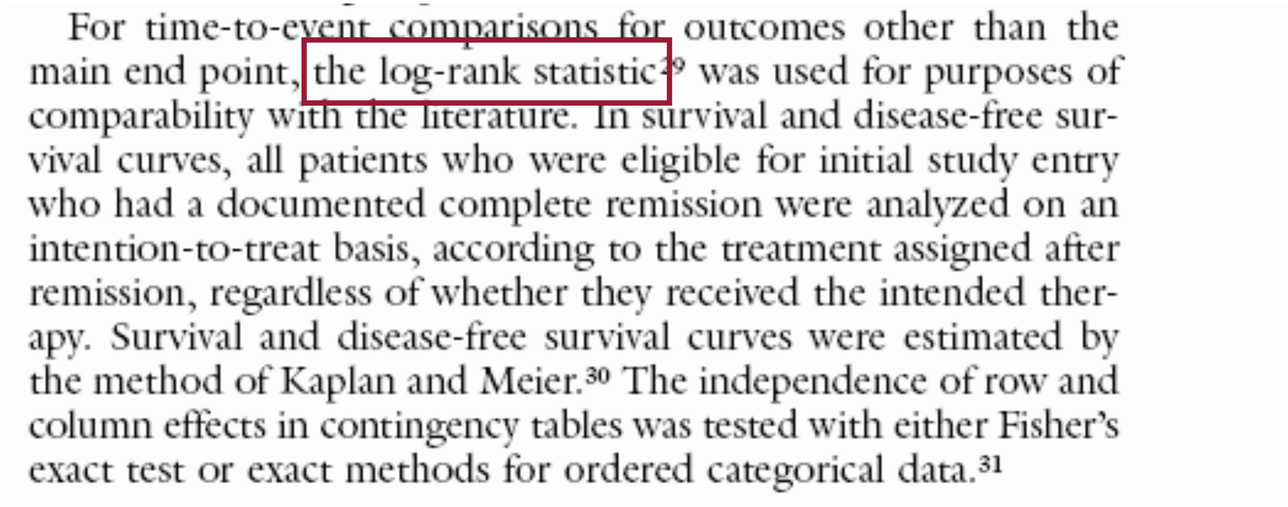
Group	No. of Events/No. at Risk				
Autologous transplantation	27/116	27/87	5/56	3/43	0/30
Allogeneic transplantation	38/113	9/74	8/61	2/36	0/25
Cytarabine	12/117	30/104	11/72	1/47	1/29

Figure 2. Probability of Survival According to Postremission Therapy.

On page 1649

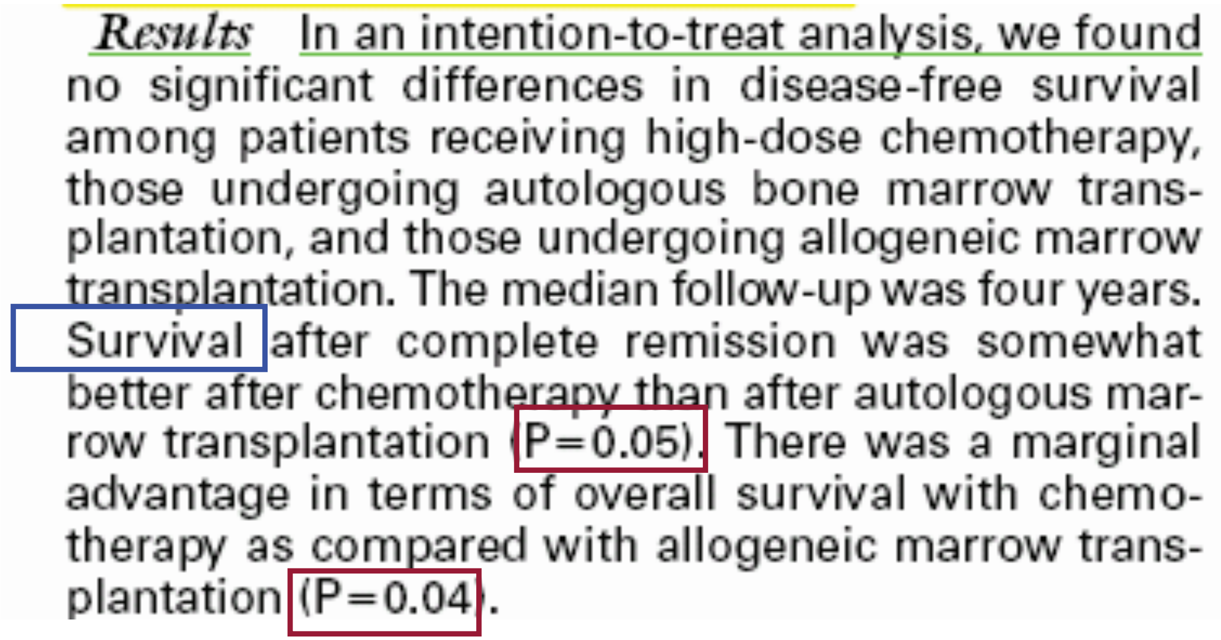


Gage cure-rate model.[26] Accrual and follow-up goals were set to provide the study with at least 80 percent power to detect a 50 percent increase in the cure rate and a 50 percent increase in median disease-free survival among the patients destined to relapse, with the use of a generalized Wilcoxon 5 percent two-sided test.[27] The study design called for a total of approximately 130 patients randomly assigned to each of the two therapies — autologous marrow transplantation and high-dose cytarabine — with 180 relapses expected in order to achieve the desired power. The protocol pro-



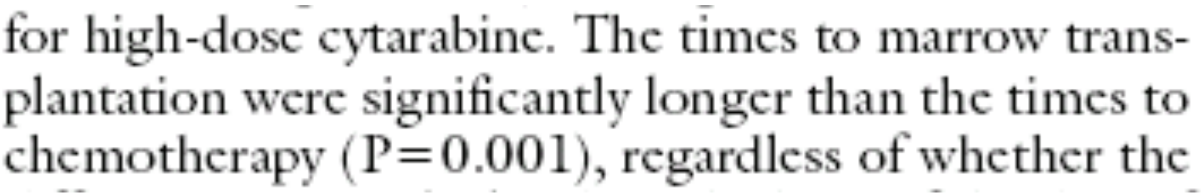
For time-to-event comparisons for outcomes other than the main end point, the log-rank statistic[29] was used for purposes of comparability with the literature. In survival and disease-free survival curves, all patients who were eligible for initial study entry who had a documented complete remission were analyzed on an intention-to-treat basis, according to the treatment assigned after remission, regardless of whether they received the intended therapy. Survival and disease-free survival curves were estimated by the method of Kaplan and Meier.[30] The independence of row and column effects in contingency tables was tested with either Fisher's exact test or exact methods for ordered categorical data.[31]

From the abstract



Results In an intention-to-treat analysis, we found no significant differences in disease-free survival among patients receiving high-dose chemotherapy, those undergoing autologous bone marrow transplantation, and those undergoing allogeneic marrow transplantation. The median follow-up was four years. Survival after complete remission was somewhat better after chemotherapy than after autologous marrow transplantation (P=0.05). There was a marginal advantage in terms of overall survival with chemotherapy as compared with allogeneic marrow transplantation (P=0.04).

On page 1652



for high-dose cytarabine. The times to marrow transplantation were significantly longer than the times to chemotherapy (P=0.001), regardless of whether the

3.3.6 The Tarone-Ware class of tests

This general class of tests is like the logrank test, but adds weights w_j. The logrank test, Wilcoxon test, and Peto-Prentice Wilcoxon are included as special cases.

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

Test	**Weight** w_j
Logrank	$w_j = 1$
Gehan's Wilcoxon	$w_j = r_j$
Peto/Prentice	$w_j = n\widehat{S}(t_j)$
Fleming-Harrington	$w_j = [\hat{S}(t_j)]^p \; [1 - \hat{S}(t_j)]^q$
Tarone-Ware	$w_j = \sqrt{r_j}$

Note: these weights w_j are not the same as the scores w_j we've been talking about earlier, and they apply to the CMH-type form of the test statistic rather than $\sum(scores)$ over a single treatment group.

More details on the Fleming-Harrington test:

The parameters p and q can be any non-negative numbers:

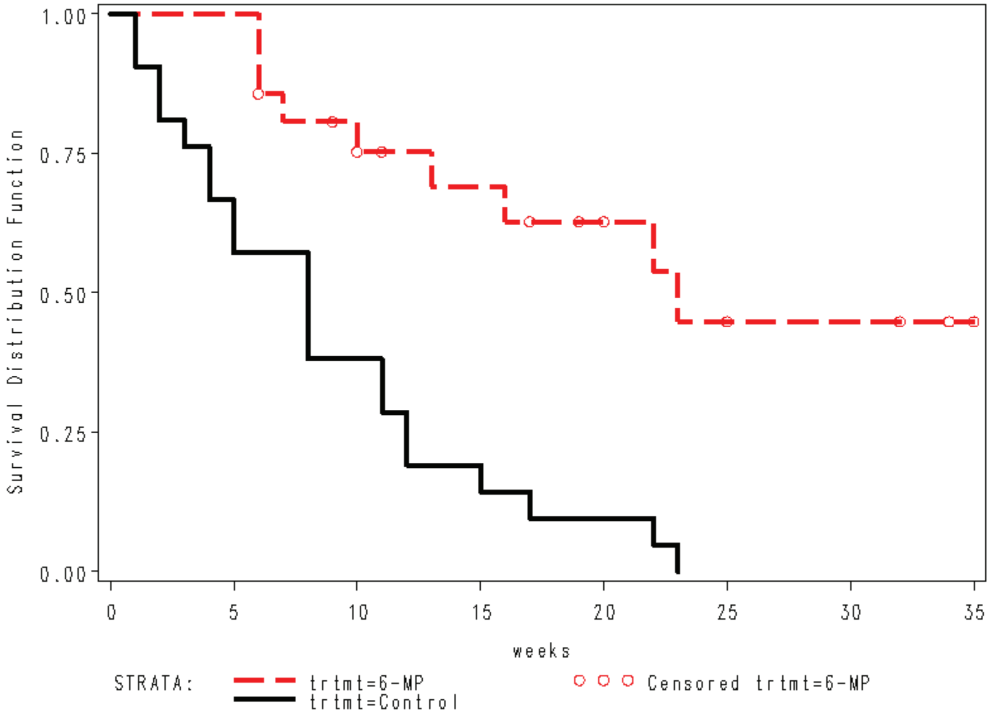
- **If p and q are both equal to 0, then $w_j = 1$ and we get the usual logrank test**
- **If $p = 1$ and $q = 0$, then the test is similar to the Peto-Prentice test (this is the default "Fleming" test in SAS PROC LIFETEST)**
- **If $q = 1$ and $p = 0$, what happens to w_j over follow-up time?**
- **If p and q are both equal to 1, the weight w_j reaches a maximum at the median, and is smaller for both large and small t_j.**

Back to our example: Time to relapse of leukemia patients

Based on the Tarone-Ware class of tests, which weights might yield the most powerful test?

Figure 3.1: Kaplan-Meier curves for both trt groups

Leukemia study (Cox and Oakes)



Obtaining Tarone-Ware class of tests in SAS:

In SAS, these two-sample comparisons can be obtained by adding the TEST=ALL option to the STRATA statement.

```
proc lifetest data=leukem outsurv=survdat;
  strata trtmt / test=all;
  time remiss*status(0);
  title 'Logrank test with proc lifetest - strata statement';
  title2 'With Tarone-Ware tests for comparing 2 samples';
run;
```

OUTPUT FROM SAS:

```
        Test of Equality over Strata

                                              Pr >
Test            Chi-Square      DF      Chi-Square

Log-Rank           16.7929       1        <.0001
Wilcoxon           13.4579       1        0.0002
Tarone             15.1236       1        0.0001
Peto               14.0841       1        0.0002
Modified Peto      13.9113       1        0.0002
Fleming(1)         14.4572       1        0.0001
```

3.3.7 Which test should we use?

CMH-type or Linear Rank?

If there are not a high proportion of ties, then it doesn't really matter since:

- **The two Wilcoxons are similar to each other**
- **The two logrank tests are similar to each other**

Logrank or Wilcoxon?

- **Both tests have the right Type I level for testing the null hypothesis of equal survival,** $H_o : S_1(t) = S_2(t)$
- **The choice of which test may therefore depend on the alternative hypothesis, which will drive the power of the test.**

- The Wilcoxon is sensitive to early differences between survival, while the logrank is sensitive to later ones. This can be seen by the relative weights they assign to the test statistic:

$$\textbf{LOGRANK} \quad \textit{numerator} = \sum_j (o_j - e_j)$$

$$\textbf{WILCOXON} \quad \textit{numerator} = \sum_j r_j (o_j - e_j)$$

- The logrank is most powerful under the assumption of proportional hazards:

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \theta$$

  which implies an alternative in terms of the survival functions of $H_a : S_1(t) = [S_2(t)]^\theta$

- The Wilcoxon has high power when the failure times are lognormally distributed, with equal variance in both groups but a different mean. It will turn out that this is the assumption of an accelerated failure time model.

- Both tests will lack power if the survival curves (or hazards) "cross". However, that does not necessarily make them *invalid*...

Comparison between TEST and STRATA in SAS for 2 examples:

Data from Lee (n=10):

from STRATA:

```
Test of Equality over Strata
                                Pr >
Test      Chi-Square    DF  Chi-Square

Log-Rank      6.9540     1      0.0084
Wilcoxon      5.5479     1      0.0185 **this is Gehan's test
-2Log(LR)     3.3444     1      0.0674
```

from TEST:

```
Univariate Chi-Squares for the WILCOXON Test

               Test       Standard                     Pr >
Variable    Statistic    Deviation    Chi-Square    Chi-Square

GROUP          1.8975       0.7508        6.3882        0.0115


Univariate Chi-Squares for the LOG RANK Test

               Test       Standard                     Pr >
Variable    Statistic    Deviation    Chi-Square    Chi-Square

GROUP          2.7500       1.0897        6.3684        0.0116
```

Previous example with leukemia data:

from STRATA:

```
Test of Equality over Strata

                                Pr >
Test       Chi-Square    DF   Chi-Square

Log-Rank      16.7929     1       0.0001
Wilcoxon      13.4579     1       0.0002
-2Log(LR)     16.4852     1       0.0001
```

from TEST:

```
Univariate Chi-Squares for the WILCOXON Test

               Test        Standard                    Pr >
Variable    Statistic     Deviation    Chi-Square    Chi-Square

GROUP          6.6928        1.7874       14.0216        0.0002

Univariate Chi-Squares for the LOG RANK Test

               Test        Standard                    Pr >
Variable    Statistic     Deviation    Chi-Square    Chi-Square

GROUP         10.2505        2.5682       15.9305        0.0001
```

3.4 P-sample and stratified logrank tests

We have been discussing two sample problems. In practice, more complex settings often arise:

- **There are more than two treatments or groups, and the question of interest is whether the groups differ from each other:**
  $H_0 : S_1(t) = S_2(t) = ... = S_P(t)$ **for all** t **versus** H_A : **not** H_0.

- **We are interested in a comparison between two groups, but we wish to adjust for another factor that may confound the analysis**

- **We want to adjust for lots of covariates.**

We will first talk about comparing the survival distributions between more than 2 groups, and then about adjusting for other covariates.

3.4.1 P-sample logrank

Suppose we observe data from P different groups, and the data from group p $(p = 1, ..., P)$ are:

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

We now construct a $(P \times 2)$ table at each of the K distinct death times, and compare the death rates between the P groups, conditional on the number at risk. We then combine tables using the CMH approach.

Let $t_1, ...t_K$ represent the K ordered, distinct death times.
At the j-th death time, we have the following table:

	Die/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 death time, and r_{pj} is the number at risk at that time.

If we were just focusing on this one table, then a $\chi^2_{(P-1)}$ test statistic could be constructed using “o”s and “e”s, like before.

Example: Toxicity in a clinical trial with 3 treatments

```
TABLE OF GROUP BY TOXICITY
GROUP     TOXICITY

Frequency|
Row Pct  |no      |yes     |  Total
---------+--------+--------+
       1 |     42 |      8 |     50
         |  84.00 |  16.00 |
---------+--------+--------+
       2 |     48 |      2 |     50
         |  96.00 |   4.00 |
---------+--------+--------+
       3 |     38 |     12 |     50
         |  76.00 |  24.00 |
---------+--------+--------+
Total         128       22      150

STATISTICS FOR TABLE OF GROUP BY TOXICITY

Statistic                     DF      Value        Prob
------------------------------------------------------
Chi-Square                     2      8.097       0.017
Likelihood Ratio Chi-Square    2      9.196       0.010
Mantel-Haenszel Chi-Square     1      1.270       0.260

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF     Value      Prob
----------------------------------------------------------
   1        Nonzero Correlation         1     1.270     0.260
   2        Row Mean Scores Differ      2     8.043     0.018
   3        General Association         2     8.043     0.018
```

Formal Calculations:

Let $\mathbf{O}_j = (d_{1j}, ...d_{(P-1)j})^T$ be a vector of the observed number of failures in groups 1 to $(P-1)$, respectively, at the j-th death time. Given the risk set sizes r_{1j}, ... r_{Pj}, and the fact that there are d_j deaths, then $\mathbf{O}_j$ has a distribution like a multivariate version of the Hypergeometric. $\mathbf{O}_j$ has mean:

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

and variance covariance matrix:

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

where the ℓ-th diagonal element is:

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

and the ℓm-th off-diagonal element is:

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

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

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

Generalizing to K tables

Analogous to what we did for the two sample logrank, we replace the $\mathbf{O}_j$, $\mathbf{E}_j$ and $\mathbf{V}_j$ with the sums over the K distinct death times. That is, let $\mathbf{O} = \sum_{j=1}^{k} \mathbf{O}_j$, $\mathbf{E} = \sum_{j=1}^{k} \mathbf{E}_j$, and $\mathbf{V} = \sum_{j=1}^{k} \mathbf{V}_j$. Then, the test statistic is:

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

Example:

Time taken to finish a test with 3 different noise distractions. All tests were stopped after 12 minutes.

Noise Level		
Group 1	Group 2	Group 3
9.0	10.0	12.0
9.5	12.0	12^+
9.0	12^+	12^+
8.5	11.0	12^+
10.0	12.0	12^+
10.5	10.5	12^+

Let's start the calculations ...

Observed data table:

Ordered	Group 1		Group 2		Group 3		Combined	
Times	d_{1j}	r_{1j}	d_{2j}	r_{2j}	d_{3j}	r_{3j}	d_j	r_j
8.5	1	6	0	6	0	6		
9.0	2	5	0	6	0	6		
9.5	1	3	0	6	0	6		
10.0	1	2	1	6	0	6		
10.5	1	1	1	5	0	6		
11.0	0	0	1	4	0	6		
12.0	0	0	2	3	1	6		

Expected table:

Ordered	Group 1		Group 2		Group 3		Combined	
Times	o_{1j}	e_{1j}	o_{2j}	e_{2j}	o_{3j}	e_{3j}	o_j	e_j
8.5								
9.0								
9.5								
10.0								
10.5								
11.0								
12.0								

Doing the P-sample test by hand is cumbersome ...

SAS program for *P*-sample logrank

```
Title 'Testing with noise example';
data noise;
   input testtime finish group;
   cards;
9        1    1
9.5      1    1
9.0      1    1
8.5      1    1
10       1    1
10.5     1    1
10.0     1    2
12       1    2
12       0    2
11       1    2
12       1    2
10.5     1    2
12       1    3
12       0    3
12       0    3
12       0    3
12       0    3
12       0    3 ;
proc lifetest data=noise;
  time testtime*finish(0);
  strata group; run;
```

```
         Testing Homogeneity of Survival Curves over Strata

Time Variable TESTTIME
                       Rank Statistics

            GROUP        Log-Rank     Wilcoxon

            1              4.4261       68.000
            2              0.4703       -5.000
            3             -4.8964      -63.000

      Covariance Matrix for the Log-Rank Statistics

    GROUP                1             2             3

    1              1.13644      -0.56191      -0.57454
    2             -0.56191       2.52446      -1.96255
    3             -0.57454      -1.96255       2.53709

      Covariance Matrix for the Wilcoxon Statistics

    GROUP                1             2             3

    1              284.808      -141.495      -143.313
    2             -141.495       466.502      -325.007
    3             -143.313      -325.007       468.320

               Test of Equality over Strata
                                         Pr >
          Test      Chi-Square    DF  Chi-Square

          Log-Rank     20.3844     2      0.0001
          Wilcoxon     18.3265     2      0.0001
          -2Log(LR)     5.5470     2      0.0624
```

Note: do not use Test **in SAS PROC LIFETEST if you want a** P**-sample logrank.** Test **will interpret the group variable as a measured covariate (i.e., either ordinal or continuous).**

In other words, you will get a *trend* test with only 1 degree of freedom, rather than a P-sample test with (p-1) df. For example:

```
proc lifetest data=noise;
  time testtime*finish(0);
  test group;
run;
```

SAS OUTPUT:

```
Univariate Chi-Squares for the LOG RANK Test
               Test       Standard                     Pr >
Variable    Statistic    Deviation    Chi-Square    Chi-Square

GROUP          9.3224       2.2846      16.6503       0.0001

Covariance Matrix for the LOG RANK Statistics
Variable         GROUP

GROUP          5.21957

Forward Stepwise Sequence of Chi-Squares for the LOG RANK Test
                                  Pr >      Chi-Square      Pr >
Variable     DF   Chi-Square   Chi-Square    Increment   Increment

GROUP         1     16.6503      0.0001       16.6503      0.0001
```

3.4.2 The Stratified Logrank

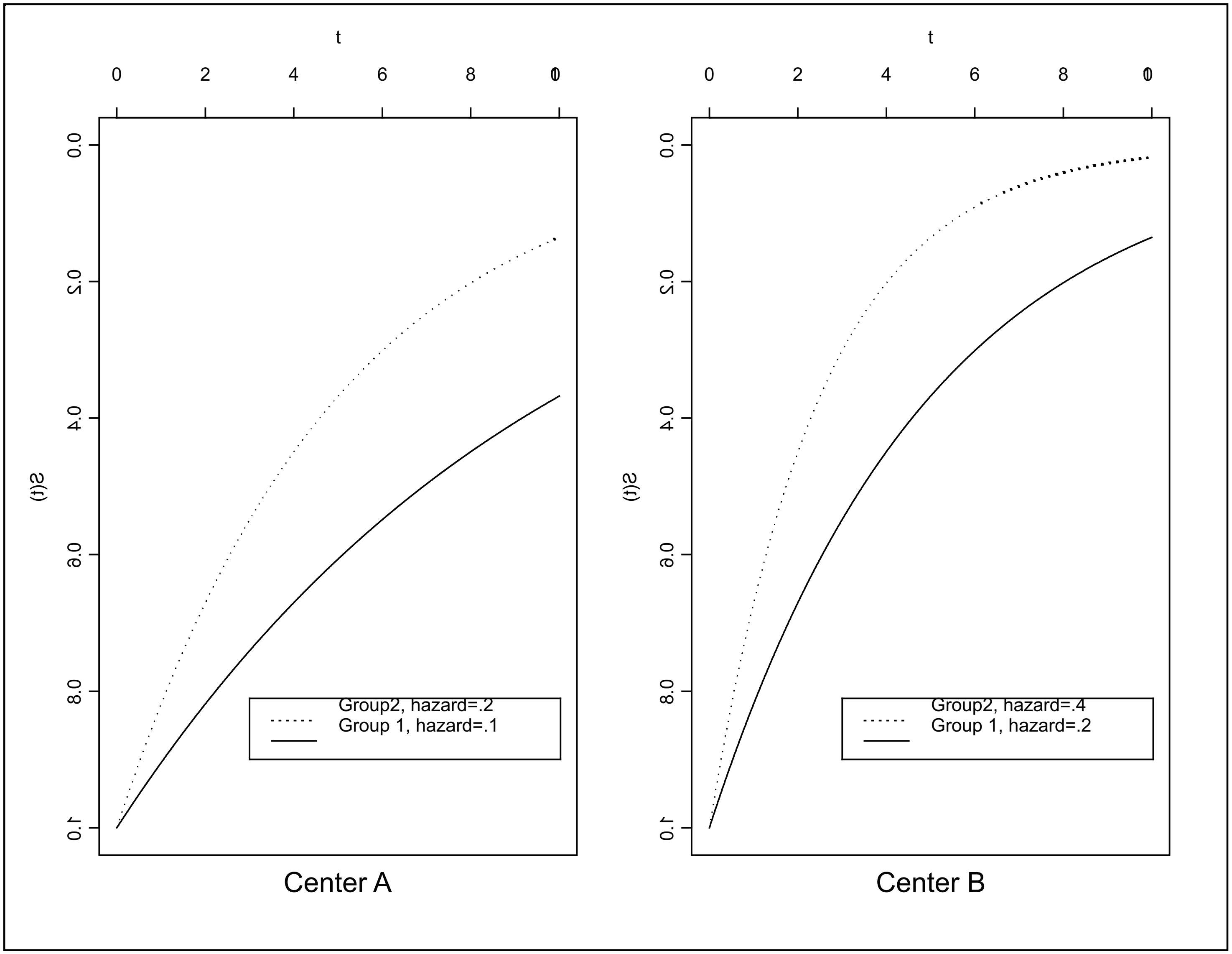
Sometimes, even though we are interested in comparing two groups (or maybe P) groups, we know there are other factors (e.g. center) that also affect the outcome. It would be useful to adjust for these other factors in some way.

Example: For the nursing home data, a logrank test comparing length of stay for those under and over 85 years of age suggests a significant difference (p=0.03).

However, we know that gender has a strong association with length of stay, and also age. Hence, it would be a good idea to STRATIFY the analysis by gender when trying to assess the age effect.

A stratified logrank allows one to efficiently compare groups, when the shapes of the hazards of the different groups differ across strata. It is most efficient when the ratio of the group 1 vs group 2 hazard is constant across strata.

In other words: $\frac{\lambda_{1s}(t)}{\lambda_{2s}(t)} = \theta$ where θ is constant over the strata $(s = 1, ..., S)$.


t
0
2
4
6
8
10
S(t)
0.0
0.2
0.4
0.6
0.8
1.0
Group2, hazard=.2
Group 1, hazard=.1
Center A
t
0
2
4
6
8
10
S(t)
0.0
0.2
0.4
0.6
0.8
1.0
Group2, hazard=.4
Group 1, hazard=.2
Center B

General setup for the stratified logrank

Suppose we want to assess the association between survival and a factor (call this X) that has two different levels. Suppose however, that we want to stratify by a second factor, that has S different levels.

First, divide the data into S separate groups. Within group s ($s = 1, ..., S$), proceed as though you were constructing the logrank to assess the association between survival and the variable X. That is, let $t_{1s}, ..., t_{K_s s}$ represent the K_s ordered, distinct death times in the s-th group.

At the j-th death time in group s, we have the following table:

	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}

Let O_s be the sum of the “o”s obtained by applying the logrank calculations in the usual way to the data from group s. Similarly, let E_s be the sum of the “e”s, and V_s be the sum of the “v”s.

The stratified logrank is

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

Note how the expected values are caculated based on data from their own strata only

When the statements ‘strata’ and ‘test’ are used jointly in the SAS lifetest procedure, then we will get the logrank test for the comparison of the ‘test variable’ while adjusting by stratification for the ‘strata‘ variable.

Stratified logrank using SAS:

```
data pop1;
  set pop;
  age1=0;
  if age >85 then age1=1;

proc lifetest data=pop1 outsurv=survres;
  time stay*censor(1);
  test age1;
  strata gender;
```

```
The LIFETEST Procedure

Rank Tests for the Association of LSTAY with Covariates
Pooled over Strata

             Univariate Chi-Squares for the LOG RANK Test
                  Test       Standard                      Pr >
    Variable    Statistic    Deviation    Chi-Square     Chi-Square

    AGE1          29.1508      17.1941       2.8744        0.0900

            Covariance Matrix for the LOG RANK Statistics

                               Variable          AGE1

                               AGE1           295.636

      Forward Stepwise Sequence of Chi-Squares for the LOG RANK Test
                                      Pr >        Chi-Square        Pr >
Variable      DF    Chi-Square     Chi-Square      Increment      Increment

AGE1           1       2.8744        0.0900          2.8744        0.0900
```

Contents