STAT331

Logrank Test

Introduction: The logrank test is the most commonly-used statistical test for comparing the survival distributions of two or more groups (such as different treatment groups in a clinical trial). The purpose of this unit is to introduce the logrank test from a heuristic perspective and to discuss popular extensions. Formal investigation of the properties of the logrank test will be covered in later units.

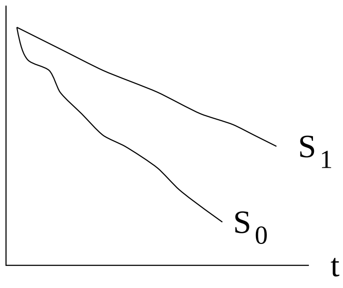
Assume that we have 2 groups of individuals, say group 0 and group 1. In group j, there are n_j i.i.d. underlying survival times with common c.d.f. denoted $F_j(\cdot)$, for j=0,1. The corresponding hazard and survival functions for group j are denoted $h_j(\cdot)$ and $S_j(\cdot)$, respectively.

As usual, we assume that the observations are subject to noninformative right censoring: within each group, the T_i and C_i are independent.

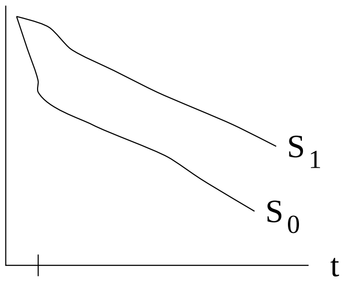
We want a nonparametric test of $\quad H_0 \; : \; F_0(\cdot) = F_1(\cdot), \quad$ or equivalently, of $S_0(\cdot) = S_1(\cdot), \quad$ or $\quad h_0(\cdot) = h_1(\cdot)$.

If we knew F_0 and F_1 were in the same parametric family (e.g., $S_j(t) = e^{-\lambda_j t}$), then H_0 is expressible as a point/region in a Euclidean parameter space. However, we instead want a nonparametric test; that is, a test whose validity does not depend on any parametric assumptions.

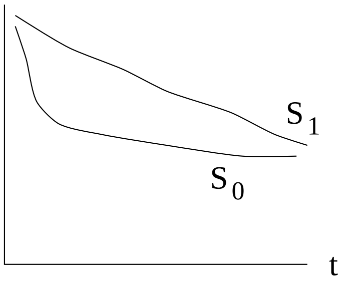
As the following picture shows, there are many ways in which $S_0(\cdot)$ and $S_1(\cdot)$ can differ:



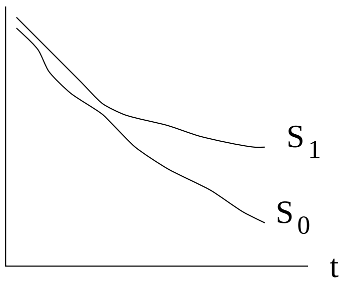
diverging



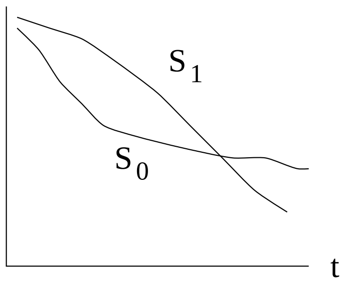
parallel after t$_0$



transient difference



late emerging difference



not stochastically ordered

It is intuitively clear that a UMP (Uniformly Most Powerful) test cannot exist for

$$\begin{array}{rrcl} & H_0 & : & S_0(\cdot) = S_1(\cdot) \\ \text{vs} & H_1 & : & \text{not } H_0 \end{array}$$

Two options in this case are to select a directional test or an omnibus test.

(a) **directional test:** These are oriented to a specific type of difference; e.g., $S_1(t) = [S_0(t)]^\theta$ for some θ. As a result, they might (and often do) have poor power against certain other alternatives.

(b) **omnibus test:** These tests attempt to have some power against most or all types of differences. As a result, they sometimes have substantially lower power than a directional test for certain alternatives. For example, a test might be based on $\int \mid \hat{S}_1(t) - \hat{S}_2(t) \mid dt$ over some time interval.

It is difficult to make the choice between directional tests, or between directional vs omnibus tests, in the abstract. It involves several factors, including prior expectations of the likely differences, properties of various tests for a variety of settings, and practical consequences of a false negative result.

Logrank Test: Early work (1960s) in this area fell along 2 lines:

(a) Modify rank tests to allow censoring (Gehan, 1965).

(b) Adapt methods used for analyzing 2×2 contingency tables to accommodate censoring (Mantel, 1966).

We introduce the logrank test from the latter perspective as it easily includes tests developed from the former and provides good insight into the properties of the logrank test.

Logrank Test Construction: Denote the distinct times of observed failures as $\tau_1 < \tau_2 < \cdots < \tau_k$, and define

$$\begin{aligned}
Y_i(\tau_j) &= \# \text{ persons in group } i \text{ who are at risk at } \tau_j \quad (i = 0, 1; j = 1, 2, \ldots, k) \\
Y(\tau_j) &= Y_0(\tau_j) + Y_1(\tau_j) = \# \quad \text{at risk at } \tau_j \text{ (both groups)} \\
d_{ij} &= \# \text{ in group } i \text{ who fail (uncensored) at } \tau_j \quad (i = 0, 1; j = 1, 2, \ldots, k) \\
d_j &= d_{0j} + d_{1j} = \quad \text{total} \quad \# \text{ failures at } \tau_j
\end{aligned}$$

The information at time τ_j can be summarized in the following 2x2 table:

	observed to fail at τ_j		at risk at τ_j
group 0	d_{0j}	$Y_0(\tau_j) - d_{0j}$	$Y_0(\tau_j)$
group 1	d_{1j}	$Y_1(\tau_j) - d_{1j}$	$Y_1(\tau_j)$
	d_j	$Y(\tau_j) - d_j$	$Y(\tau_j)$

Note: $d_{0j}/Y_0(\tau_j)$ can be viewed as an estimator of $h_0(\tau_j)$.

Suppose $H_0 : F_0(\cdot) = F_1(\cdot)$ holds. Conditional on the 4 marginal totals, a single element (say d_{1j}) defines the table. Furthermore, with this conditioning and assuming H_0, d_{1j} has the hypergeometric distribution; that is:

$$P[d_{1j} = d] = \left(\begin{array}{c} d_j \\ d \end{array} \right) \left(\begin{array}{c} Y(\tau_j) - d_j \\ Y_1(\tau_j) - d \end{array} \right) \Big/ \left(\begin{array}{c} Y(\tau_j) \\ Y_1(\tau_j) \end{array} \right) \quad \text{for}$$
$$d = max(0, d_j - Y_0(\tau_j)), \cdots, min(d_j, Y_1(\tau_j)).$$

The mean and variance of d_{1j} under H_0 are thus

$$\begin{aligned} E_j &= \left(\tfrac{Y_1(\tau_j)}{Y(\tau_j)}\right) d_j \\ V_j &= \tfrac{Y(\tau_j) - Y_1(\tau_j)}{Y(\tau_j) - 1} \ \cdot \ Y_1(\tau_j) \left(\tfrac{d_j}{Y(\tau_j)}\right) \left(1 - \tfrac{d_j}{Y(\tau_j)}\right) \\ &= \tfrac{Y_0(\tau_j) Y_1(\tau_j) d_j (Y(\tau_j) - d_j)}{Y(\tau_j)^2 (Y(\tau_j) - 1)} \end{aligned}$$

Define $O_j = d_{1j}$. Fisher's test would tell us to consider extreme values of d_{1j} as evidence against H_0.

Thus, define

$$\begin{aligned} O &= \textstyle\sum_{j=1}^{k} O_j = \text{ total \# failures in group 1} \\ E &= \textstyle\sum_{j=1}^{k} E_j \\ V &= \textstyle\sum_{j=1}^{k} V_j \end{aligned}$$

and let

$$Z = \frac{O - E}{\sqrt{V}} = \frac{\displaystyle\sum_j (O_j - E_j)}{\sqrt{\displaystyle\sum_j V_j}}.$$

Then under H_0, it is argued that

$$Z \overset{\text{apx}}{\sim} N(0, 1)$$

$$(\text{or that } Z^2 \overset{\text{apx}}{\sim} \chi^2_1)$$

This approximation can be used to obtain an approximate test for H_0 by comparing the observed value of Z (or Z^2) to the tail area of the standard normal (chi-square) distribution.

Example:

$$\begin{aligned} \text{Group 0} &: 3.1, 6.8^+, 9, 9, 11.3^+, 16.2 \\ \text{Group 1} &: 8.7, 9, 10.1^+, 12.1^+, 18.7, 23.1^+ \end{aligned}$$

Then $k = 5$ and $\tau_1, \ldots, \tau_5 = 3.1, 8.7, 9, 16.2, 18.7$

	$\tau_1 = 3.1$			$\tau_2 = 8.7$			$\tau_3 = 9$			$\tau_4 = 16.2$			$\tau_5 = 18.7$		
Group 0	1	5	6	0	4	4	2	2	4	1	0	1	0	0	0
Group 1	0	6	6	1	5	6	1	4	5	0	2	2	1	1	2
	1	11	12	1	9	10	3	6	9	1	2	3	1	1	2

$O_j =$	0	1	1	0	1
$E_j =$	1/2	6/10	15/9	2/3	1
$V_j =$	1/4	6/25	5/9	2/9	0

$$O = 3, \quad E = 3.44, \quad V = 1.26, \quad Z = -.39 \quad (\text{2-sided } P = .70)$$

Comments:

- Note that the test statistics is only affected by ranks of the observed times (both censored or failure).

- While E_j may be a conditional expectation for each j, it is not clear that E has such an interpretation. Also, the creation of Z and its approximation as a $N(0, 1)$ r.v. suggests that the contributions from each τ_j are independent. Is this true/accurate? Then, is $Z \xrightarrow{\mathcal{L}} N(0, 1)$ under H_0?

- Note the similarity of the logrank test to techniques for combining 2×2 tables across strata (e.g., cities).

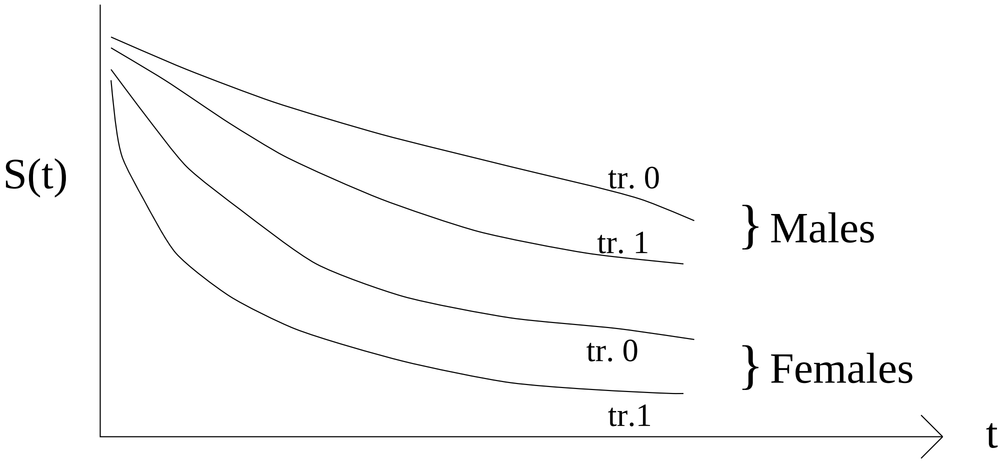
- Note that the sequences $Y_0(\tau_1), Y_0(\tau_2), Y_0(\tau_3), \ldots$ and $Y_1(\tau_1), Y_1(\tau_2), Y_1(\tau_3), \ldots$ are nonincreasing, and as soon as one reaches 0 [e.g., $Y_0(\tau_5) = 0$ at

$\tau_5 = 18.7$], it must follow that $O_j = E_j$ and $V_j = 0$ at and beyond this time. Thus, we would get the same answer (i.e., Z) if the construction stopped at the last time when both $Y_0(\tau_j)$ and $Y_1(\tau_j)$ are > 0.

- Although it is not obvious from the construction, the logrank test is a directional test oriented towards alternatives where $S_1(t) = (S_0(t))^\theta$, or equivalently, when $h_1(t)/h_0(t) = \theta$. We will see later that the logrank statistic arises as a score test from a partial likelihood function for Cox's proportional hazards model.
- While the heuristic arguments leading to the approximation of the null distribution of the logrank test seem reasonable, is the result correct? In addition, how does the test behave as a function of the amount of censoring or the hazard functions in the two treatment groups? We return to these important practical questions in later units.

3 Some Extensions of the Logrank Test:

Stratified logrank test: Suppose that we have two groups (say, 2 treatments), as before, but that we want to control (adjust) for a categorical covariate (e.g., gender). Then there are 4=2x2 types of individuals. For example, their respective survivor functions might be as shown below. If we still want to compare treatment groups, but also 'adjust' for gender, a stratified logrank test could be used. Suppose $S_j^{(l)}(\cdot)$ denotes the survival function for group j in stratum l, and consider $H_0 : S_0^{(l)}(\cdot) = S_1^{(l)}(\cdot)$, $l = 1, \cdots, L$.


S(t)
tr. 0
tr. 1
} Males
tr. 0
} Females
tr.1
t

The stratified logrank test is useful when the distibution of the stratum variable in the two groups is not the same, but the distribution of the relevant covariates in each stratum is the same in both groups (within each stratum, the groups have a comparable prognosis). The stratified logrank test can also be useful to gain precision.

Construction

1. Separate data into L groups, where $L = \#$ levels of the categorical covariates on which you want to stratify (e.g., $L = 2$ when stratifying by gender)

2. Compute $O,\ E,\ V$ (say, $O^{(l)},\ E^{(l)},\ V^{(l)}$) within each group, just as with the ordinary logrank

3. $Z = \dfrac{\displaystyle\sum_{l=1}^{L}(O^{(l)} - E^{(l)})}{\sqrt{\displaystyle\sum_{l=1}^{L} V^{(l)}}} \overset{\text{apx}}{\approx} N(0,1)$ under H_0

Note 1: Intuitively, it should be clear how this statistic attempts to adjust for the stratification variable and, assuming the $[O^{(l)}, E^{(l)}, V^{(l)}]$ are approximately uncorrelated, that the statistic will be approximately $N(0,1)$.

Note 2: If there are too many strata, the test could have poor power. In part, this would be due to the feature of the logrank test that there is no contribution for any 2x2 table once one of the $Y_l(\tau_j)$ becomes zero.

Note 3: As we will later see, the stratified logrank test also arises as a score test from Cox's model. This relationship will also clarify the types of alternatives to H_0 for which the stratified logrank test is directed.

Weighted Logrank Test: Note that in the logrank test, $O_j - E_j$ is a measure of how $h_0(\tau_j)$ and $h_1(\tau_j)$ differ.

Suppose we wanted to compare groups, but in a way that 'emphasized' certain times more than others.

Let $w_1 \geq 0,\ w_2 \geq 0, \ldots,\ w_K \geq 0$ be known constants. Then the weighted logrank test is given by

$$Z_w = \frac{\displaystyle\sum_{j=1}^{K} w_j(O_j - E_j)}{\sqrt{\displaystyle\sum_{j=1}^{K} w_j^2 V_j}}$$

and, under H_0, $Z_w \overset{\text{apx}}{\approx} N(0,1)$.

Note:

- Choosing $W_j = w$ (i.e., constant in j) yields the ordinary logrank test.

- Perhaps choose larger weights for those τ_j where a larger difference is anticipated. But what does "difference" refer to?

  $h_0(\tau_j) - h_1(\tau_j),\ h_0(\tau_j)/h_1(\tau_j),\ \ S_0(\tau_j)/S_1(\tau_j)$??? (more later).

- Special case where $w_j = Y(\tau_j)$ yields what is sometimes called the Generalized Wilcoxon test.

- Since $Y(\tau_1) > Y(\tau_2) > Y(\tau_3) > \cdots$, the Generalized Wilcoxon test places (relatively) greater emphasis on early differences between $h_0(\cdot)$ and $h_1(\cdot)$ than the logrank test.

- Suppose that two-sample data consists of $\{(U_{ij}, \delta_{ij}), j = 1, \cdots, n_i, i = 0, 1\}$, then the test statistics for the generalized Wilcoxon test before standardization can be written as

$$\sum_{i=1}^{n_1}\sum_{j=1}^{n_0}\{I(U_{1i} \geq U_{0j})\delta_{0j} - I(U_{0j} \geq U_{1i})\delta_{1i}\}$$

which becomes

$$2\sum_{i=1}^{n_1}\sum_{j=1}^{n_0}\left\{I(U_{1i} > U_{0j}) - \frac{1}{2}\right\}$$

the commonly used Wilcoxon test statistics in the absence of censoring.

Example (revisited): Using $W_j = Y(\tau_j)$ yields $Z_w = -.97$
(2-sided $P = .33$)

Several questions arise from these considerations:

- Is the weighted logrank asymptotically $N(0,1)$ under H_0?

- The weights used above (i.e., $W_j = Y(\tau_j)$) are data dependent (that is, r.v.'s). How does this impact the asymptotic behavior of the test statistic?

- How does one pick the W_j?

  We will return to these issues in a later unit.

Logrank Test for $>$ 2 Groups: Now suppose that we wish to compare the survival distributions of several (>2) groups. Specifically, suppose there are $p+1$ groups, denoted $0, 1, 2, \ldots, p$, and that we wish to test the hypothesis:

$$H_0: \; S_0(\cdot) = S_1(\cdot) = \cdots = S_p(\cdot)$$

$$\begin{array}{lrcl} \text{e.g.} & \text{Group } 0 & = & \text{placebo group} \\ & \text{Group } j & = & \text{dose } D_j \text{ of a drug} \qquad j = 1, 2, \ldots, p \\ & & & (D_1 < D_2 < \cdots < D_p) \end{array}$$

Then an extension of the usual logrank test (where $p = 1$) for this setting is given as follows (assume H_0):

Construction

	fail at τ_j		at risk at τ_j
Group 0	d_{0j}	$Y_0(\tau_j) - d_{0j}$	$Y_0(\tau_j)$
Group 1	d_{1j}	$Y_1(\tau_j) - d_{1j}$	$Y_1(\tau_j)$
Group 2	d_{2j}	$Y_2(\tau_j) - d_{2j}$	$Y_2(\tau_j)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group p	d_{pj}	$Y_p(\tau_j) - d_{pj}$	$Y_p(\tau_j)$
Total	d_j	$Y(\tau_j) - d_j$	$Y(\tau_j)$

$$\mathbf{0}_j \overset{p \times 1}{=} \begin{pmatrix} d_{1j} \\ d_{2j} \\ \vdots \\ d_{pj} \end{pmatrix}, \quad \mathbf{E}_j \overset{p \times 1}{=} \begin{pmatrix} E_{1j} \\ E_{2j} \\ \vdots \\ E_{pj} \end{pmatrix} \quad \text{where } E_{ij} = \frac{Y_i(\tau_j)}{Y(\tau_j)} \cdot d_j$$

$$\mathbf{V_j} \overset{p \times p}{=} \left(V_{kl}^{(j)}\right), \quad \text{where}$$

$$V_{kl}^{(j)} = \frac{d_j Y_k(\tau_j)(Y(\tau_j) - d_j)(Y(\tau_j) \cdot 1(k = l) - Y_k(\tau_j))}{Y(\tau_j)^2 (Y(\tau_j) - 1)}$$

Then with

$$\begin{aligned}\mathbf{O}_{.} &= \sum_{j=1}^{k} \mathbf{O}_j \\ \mathbf{E}_{.} &= \sum_{j=1}^{k} \mathbf{E}_j \\ \mathbf{V} &= \sum_{j=1}^{k} \mathbf{V_j},\end{aligned}$$

$$Q_p = (\mathbf{O}_{.} - \mathbf{E}_{.})^T \, \mathbf{V}^{-1} \, (\mathbf{O}. - \mathbf{E}.) \;\overset{\text{apx}}{\sim}\; \chi^2_p \text{ under } H_0$$

Note: This test is 'omnibus' in terms of how it combines the $p+1$ groups; i.e., it is not directed towards a dose-response, in contrast to the trend test below.

Logrank trend test: How can we modify this to test for a trend in the survival functions / hazard functions in the $p+1$ groups? For example, suppose the groups have a natural ordering, such as increasing exposures to a toxic substance or increasing doses of a drug. Then one might expect the risk of failure to be monotone with exposure/dose and thus want to design a test that is especially oriented towards this type of alternative to H_0.

Let $\mathbf{c}$ = any $p \times 1$ vector of constants. If $\mathbf{O}_{.} - \mathbf{E}_{.} \overset{\text{apx}}{\sim} N(\mathbf{0}, \mathbf{V})$ under H_0,

↖ zero vector

$$\mathbf{c}^T(\mathbf{O}_{.} - \mathbf{E}_{.}) \;\overset{\text{apx}}{\sim}\; N(0, \mathbf{c}^T\mathbf{V}\mathbf{c})$$

↪ Trend test

$$Z_{tr} = \frac{\mathbf{c}^T(\mathbf{O}_{.} - \mathbf{E}_{.})}{\sqrt{\mathbf{c}^T\mathbf{V}\mathbf{c}}} \;\overset{\text{apx}}{\sim}\; N(0,1) \text{ under } H_0$$

e.g., take $c_j = D_j$ (dose used for group j), $j = 1, 2, \ldots, p$.

Using the relation of the (weighted) logrank test to the Cox partial likelihood approach, one can test whether the trend is the only cause of variation between the p groups. Comparing the partial log likelihood of the model with the trend variable and the model with the trend variable and $p-1$ dummies can be used as a check for deviation from the directional alternative assumed by Z_{tr} (this is the same as comparing the model with the trend variable and the model with p dummies). We will come back to this in a later unit.

Several questions arise from these considerations:

- How to choose $\mathbf{c}$? The choice of weights $\mathbf{c}$ depends in part on the setting. For example, for some purposes the key scientific question might be whether or not the risk is monotone with the $p+1$ groups, while in others it might be important to distinguish a linear versus supralinear (e.g., quadratic) dose-response. It may be clear that we want the components of $\mathbf{c}$ to be monotone, but against what specific alternative is a particular choice of $\mathbf{c}$ optimal and what are the consequences of selecting the 'wrong' value of $\mathbf{c}$? It will later be seen that the logrank trend test arises as a likelihood score test from Cox's proportional hazards model. This link will not only provide the basis for the asymptotic behavior of the trend test, but also clarify the implications for a particular choice of the vector $\mathbf{c}$.
- When to use Z_{tr} vs Q_p?

  We return to this in a later unit.

Finally, we note that these variations of the logrank test can be combined. For example, we can do a stratified version of the logrank test with $P > 2$ groups, a stratified trend test, etc.

SAS commands for logrank test:

Consider the dataset AZT-ddI.dat. For dataset AZTddI, with “Tad” as the time variable, “ad” as the censoring variable (ad=0 indicates censoring, ad=1 indicates event), and “rx” as the grouping variable, the following code does a logrank test and a generalized Wilcoxon test comparing the 2 levels of group:

```
proc lifetest data=AZTddI;
   time Tad *ad(0);
   strata rx / test=(logrank Wilcoxon);
run;
```

Stratified logrank test, stratified by “gender”:

```
proc lifetest data=AZTddI;
   time Tad *ad(0);
   strata gender / group=rx;
run;
```

Now suppose that we have $p + 1 > 2$ levels in the group variable, such as 3 treatments. Then the same commands can be applied but one gets the p df version of the logrank or wilcoxon test (including stratified or not stratified).

To get the trend test with (linear) weights: if the variable is numeric, the unformatted values of the variable are used as the scores; otherwise, the scores are 1, 2, ... , in the given order of the strata. For as_ar_a (categories asymptomatic/aids-related-complex/aids) use e.g.:

```
proc lifetest data=AZTddI;
   time Tad *ad(0);
   strata as_ar_a / trend;
run;
```

STATA commands for logrank test:

After reading in the data and using the **.stset** command to define U and δ:

Suppose there are 2 groups and **group** is the variable defining group.

.sts test group or **.sts test group,logrank**: Either does a logrank test comparing the 2 levels of **group**.

.sts test group,wilcoxon: This does the generalized wilcoxon test

.sts test group, logrank by(gender): This does the stratified logrank test comparing the 2 levels of **group**, with stratification by **gender**

Now suppose that we have $p + 1 > 2$ levels in the variable group, such as 3 treatments. Then the same commands can be applied but one gets the p df version of the logrank or wilcoxon test (including stratified or not stratified).

To get the trend test with (linear) weights (0,1,2,...,p), use:

.sts test group,trend

It will give both the p df test and the trend test.

R commands for logrank test:

```
survdiff(Surv(time, delta)~ group+strata(gender))
```

Exercises

1. For each of the five pairs of survival curves shown on page 2, sketch the corresponding pairs of hazard functions.

2. For the data set in freireich.dat, compute by hand the logrank test for comparing treatment groups by forming each 2x2 table and recording the entries, and then computing the components of the test statistics. Verify your answer by using STATA, SAS, or any of your favorite programs to compute the logrank test statistic.

3. Analyze the data in AZT-ddI (with U=Tad, and δ=ad). First, create a new variable called cd4cat defined as 0 if cd4$<$ 100, 1 if cd4 $\in [100, 199]$ and 2 if cd4 $\geq$ 200. Then:

   (a) do a Kaplan-Meier plot by cd4cat

   (b) do a logrank test of cd4cat = 0 vs cd4cat = 1

   (c) do a Generalized Wilcoxon test of cd4cat = 0 vs cd4cat = 1

   (d) do a stratified (by gender) logrank of cd4cat = 0 vs cd4cat = 1

   (e) do a logrank test of cd4cat=0 vs cd4cat=1 vs cd4cat=2

   (f) do a logrank trend test of cd4cat (=0,1,2)

   (g) is there evidence in the data of a departure from linear trend?

   (h) do a stratified (by gender) logrank trend test of cd4cat

4. What is the rationel behind choosing $\mathbf{c}^T(\mathbf{O}_. - \mathbf{E}_.)$ for the trend test?

5. An HIV clinical trial is in preparation comparing the treatments ABC+3TC versus TDF+FTC. The efficacy endpoint will be composite: time until virologic failure, time until HIV disease progression, and death, whichever comes first.

   (a) (5 points). How would you analyze the efficacy?

   (b) (8 points). Suppose that the difference in the effect of these treatments is expected to be higher in the first time period than at later times. How could you adapt the method of (a) to get a test with higher power?

   (c) (10 points). Suppose you want to do a separate analysis for time to first virologic failure or death, whichever comes first. What kind of analysis would you propose? How would you define the failure time, and what would you consider as censoring?

   (d) (10 points). A separate analysis will be done for time to first virologic failure. In this analysis, death is considered as censoring. What assumption does this indicate? Please comment on this.

6. (25 points). Suppose that a subgroup analysis on survival indicates that treatment effect has an opposite sign in men and women; e.g., in women the side effects outweight the benefits, e.g., due to breast cancer risk.

   (a) (10 points). What do you think will be the likely outcome of the logrank test? Explain why.

   (b) (10 points). What do you think will be the likely outcome of the logrank test stratified by gender? Explain why.

   (c) (5 points). Are type-1 errors with the above methods still ok?

Additional Reading and Comments

Tarone and Ware (1977) proposed the use of a weighted version of the logrank test, and formal evaluation of the properties of such tests was developed subsequently (see, for example, Fleming & Harrington, 1991, for a review). Catherine Hill (1981) evaluated the loss in efficiency from using a stratified logrank test. Despite the widespread use of stratified logrank tests, relatively little attention appears to have been given to issues of efficiency and robustness for this approach.

References

Collett D (1994). Modelling survival data in medical research. Texts in statistical science, Chapman and Hall, London.

Gehan EA (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. Biometrika, 52: 203-223.

Fleming T $ Harrington D (1991). Counting Processes and Survival Analysis, Wiley, New York.

Hill C (1981). Asymptotic relative efficiency of survival tests with covariates. Biometrika, 68:669-702.

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

Tarone RE & Ware JH (1977). On distribution-free tests for equality of survival distributions. Biometrika, 64:156-160.