STAT331

2-sample Tests (continued)

In the last section we saw that under a sequence of contiguous alternatives to the null, the asymptotic distribution of the weighted logrank test, Z_w, was $\mathcal{N}(\theta, 1)$, where

$$\theta = \xi/\sigma_w,$$

σ_w^2 is given by (19.5) and ξ by (19.8):

$$\sigma_w^2 = \int_0^\infty \frac{f_0(s)(1-G_0(s))(1-G_1(s))w^2(s)}{((1-p)(1-G_0(s))+p(1-G_1(s)))}ds, \tag{19.5}$$

$$\xi = \sqrt{p(1-p)}\int_0^\infty \frac{f_0(s)(1-G_0(s))(1-G_1(s))w(s)g(s)ds}{((1-p)(1-G_0(s))+p(1-G_1(s)))}. \tag{19.8}$$

Let's investigate the non-centrality parameter (NCP) θ a bit more. In the following, we will suppose that $P(Z_i = 1) = p$ for all i.

Define $p(t)$ to be the null probability that someone at risk at time t is in treatment group 1; that is,

$$p(t) = P(Z_i = 1 \mid U_i \geq t)$$

under H_0. Then we have

$$p(t) = \frac{p(1-G_1(t))}{(1-p)(1-G_0(t))+p(1-G_1(t))}\ .$$

Also, let $v(t)$ denote the density (under H_0) for observing a failure at time t; that is,

$$v(t) = (\ (1-p)(1-G_0(t))+p(1-G_1(t))\ )\ f_0(t).$$

We can then re-express ξ and σ_w^2 as

$$\xi = \cdots = \frac{1}{\sqrt{p(1-p)}} \int_0^\infty p(t)(1-p(t))\ w(t)g(t)\ v(t)\ dt$$

and

$$\sigma_w^2 = \frac{1}{p(1-p)} \int_0^\infty p(t)(1-p(t))w^2(t)v(t)\ dt.$$

Therefore, we can write the NCP of the weighted logrank (LR) test Z_w as

$$\theta = \frac{\xi}{\sigma_w} = \frac{\int_0^\infty p(t)(1-p(t))w(t)g(t)\ v(t)dt}{\sqrt{\int_0^\infty p(t)(1-p(t))w^2(t)v(t)dt}}. \tag{20.1}$$

Recall that this asymptotic NCP for the weighted LR test arises under the sequence of contiguous alternatives to H_0 given by

$$H_{A,n}: \quad \ln\left(\frac{\lambda_1(t)}{\lambda_0(t)}\right) = n^{-1/2}g(t);$$

Thus, $g(t)$ is proportional to the log hazard ratio.

Applying the Cauchy-Schwartz inequality to the numerator of (20.1), we see that

$$\left|\int\ p(t)(1-p(t))v(t)\ (w(t)g(t))dt\right|$$
$$\leq \sqrt{\int p(t)(1-p(t))v(t)(w^2(t))dt}\ \sqrt{\int p(t)(1-p(t))v(t)(g^2(t))dt}.$$

Thus, we have that

$$\mid \theta \mid \leq \sqrt{\int p(t)(1-p(t))v(t)\ g^2(t)dt}$$

for any $w(t)$, with equality when $w(\cdot) \propto g(\cdot)$. Note that the right-hand-side of this inequality is equal to the absolute value of the NCP of the weighted logrank test that uses a weight function w(t) proportional to g(t); that is, proportional to the log of the hazard ratio. Thus, we can conclude that the optimal weight in this class of tests (i.e., the one maximizing the NCP) uses weights proportional to $\ln\left(\frac{\lambda_1(t)}{\lambda_0(t)}\right)$. (Since the weighted logrank test is invariant to scale changes to the weights, any set of weights that are a scale change of the optimal set is also optimal.)

Note – In practice, we don't know $\lambda_1(t)$ or $\lambda_0(t)$, and hence $g(t)$. However, this result is helpful because:

1. It tells us the best we can do with this class of tests, and thus provides a standard against which we can compare the properties of a specific weighted logrank test for various alternatives; and
2. If we have a "sense" of the nature of $\ln\left(\frac{\lambda_1(t)}{\lambda_0(t)}\right)$ (e.g., decreasing, approximately constant, etc.) based on our knowledge of the kind of data we are analyzing, we might pick a weight $w(\cdot)$ that reflects this (e.g. $w(t) = \hat{S}(t-)$ if we believe the hazard ratio is decreasing).

We now use the result in (20.1) to address 2 questions:

1. How efficient is the logrank test compared to its parametric counterpart?
2. How efficient is the logrank test compared to the optimal weighted logrank test for a variety of alternatives to H_0?

Efficiency of Logrank Test to Parametric Test for Exponential Data

Recall that if 2 test statistics are asymptotically $\mathcal{N}(0,1)$ under some null and are asymptotically Normal with variance 1 under a sequence of contiguous alternatives to the null, then their asymptotic relative efficiency (ARE) is the square of the ratio of their NCPs. The ARE can be loosely interpreted as the relative sample sizes needed so that the tests will have equal power.

To illustrate how the (nonparametric) logrank test compares to its parametric counterpart, suppose $\lambda_0(t)$ and $\lambda_1(t)$ are constants, so that $g(t)$ is also constant. Using standard likelihood methods, we can find the NCP for the parametric score test of H_0 under the alternative $H_{A,n}: \quad ln(\frac{\lambda_1(t)}{\lambda_0(t)}) = n^{-1/2}$ $(g = 1)$. See Schoenfeld (1981) for details. Call this NCP θ_{exp} and denote the logrank's NCP by θ_{LR}. Schoenfeld (1981) shows that

$$\theta^2_{exp} = \frac{p(1-p)d_0d_1}{pd_1 + (1-p)d_0},$$

where $d_j = P(\delta_i = 1 \mid Z_i = j,\ H_0) = \int_0^\infty f_0(t)(1 - G_j(t))dt \quad j = 0, 1.$

Then the asymptotic relative efficiency (ARE) of the logrank test to the exponential parametric test is

$$\text{ARE} = \left(\frac{\theta_{LR}}{\theta_{exp}}\right)^2 = \frac{pd_1 + (1-p)d_0}{p(1-p)d_0d_1} \int_0^\infty p(t)(1-p(t))v(t)dt\ . \qquad (20.2)$$

Case 1: $G_0(\cdot) = G_1(\cdot)$. When the censoring distributions in the 2 groups are equal, $d_0 = d_1$, and $p(t) = p$, so that

$$\text{ARE} = 1\ .$$

Thus, the logrank is fully efficient relative to the parametric test when the censoring distributions in the 2 groups are equal.

Case 2: When $G_0(\cdot) \neq G_1(\cdot)$, the ARE will be ≤ 1. The amount of inefficiency will depend on $G_0(\cdot)$, $G_1(\cdot)$, p, and λ_0 (= the exponential parameter).

Let's examine this for the special case where $p = \frac{1}{2}$, $G_0(t) = 1 - e^{-\varphi_0 t}$ and $G_1(t) = 1 - e^{-\varphi_1 t}$; i.e., when censoring is also exponentially distributed. Without loss of generality, we can take $\lambda_0 = 1$, so that

$$d_j = \frac{1}{\varphi_j + 1} \quad j = 0, 1.$$

It then is easy to verify that

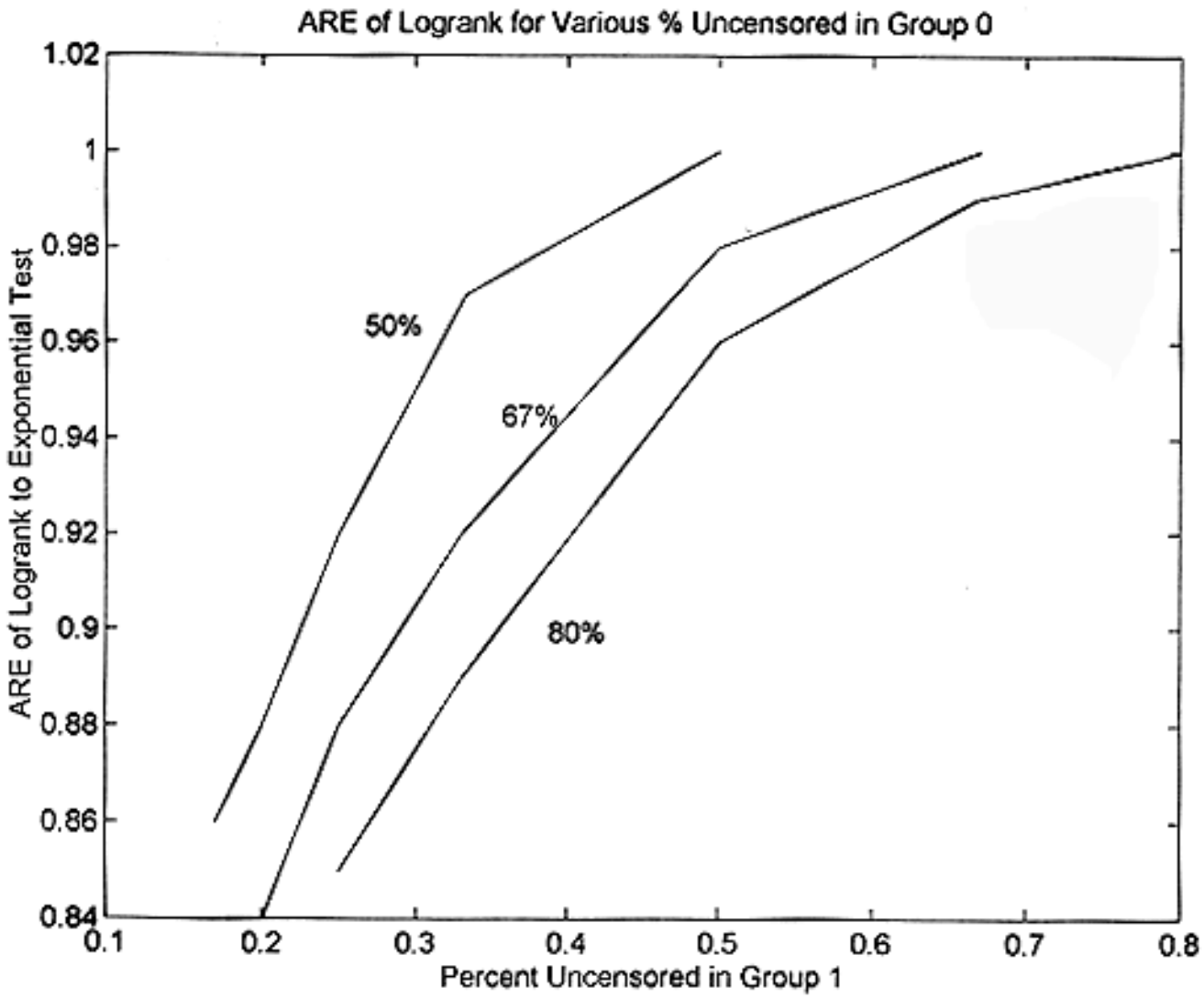
$$\text{ARE}(Z_{LR}, \text{Parametric}) = \ldots = (2 + \varphi_0 + \varphi_1) \int_0^\infty \frac{e^{-(1+\phi_0+\phi_1)t}}{e^{-\phi_0 t} + e^{-\phi_1 t}} dt.$$

Figure 20-1 gives this ARE for various choices of φ_0 and φ_1, where the latter are expressed in terms of the proportion uncensored, d_0 and d_1, in the 2 groups (note that $\varphi_j = (1 - d_j)/d_j$). For any given amount of censoring in group 0, the ARE increases with the amount of censoring in the group 1. When the censoring distributions get closer, the ARE increases. All of the AREs depicted are high (above .80).

Thus, the logrank test is fully efficient relative to the best parametric test when the 2 groups have equal censoring distributions, and still quite efficient when these censoring distributions differ.

Intuitively, it is not surprising that the logrank test would be inefficient in this setting. Recall from Unit 6 that once all sujects remaining at risk are in one group, none of the subsequent 2×2 tables used to construct the logrank test contribute to the statistic (because $O_j - E_j = 0$ and $V_j = 0$ for these tables). For a nonparametric test this is appropriate. However, a parametric test makes use of this information. For example, the exponential tests use all observed data to compute the MLE of λ for each group.

Figure 20-1


ARE of Logrank for Various % Uncensored in Group 0
ARE of Logrank to Exponential Test
Percent Uncensored in Group 1
1.02
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
50%
67%
80%

ARE of 2 Weighted Logrank Tests

The ARE of any 2 weighted logrank tests, say Z_{w_1} and Z_{w_2}, is equal to $(\theta_{w_1}/\theta_{w_2})^2$. Let's use this to examine the logrank test ($w(t) = 1$) in comparison to the optimal weighted logrank test ($w(t) = g(t)$). From (20.1),

$$\text{ARE}[Z_{LR} : \ Z_{opt}] = \frac{\left(\frac{\int p(t)(1-p(t))g(t)v(t)dt}{\sqrt{\int p(t)(1-p(t))v(t)dt}}\right)^2}{\left(\frac{\int p(t)(1-p(t))g^2(t)v(t)dt}{\sqrt{\int p(t)(1-p(t))g^2(t)v(t)dt}}\right)^2} \ ,$$

$$= \frac{\left(\int_0^\infty p(t)(1-p(t))g(t)v(t)dt\right)^2}{\left(\int_0^\infty p(t)(1-p(t))v(t)dt\right)\left(\int_0^\infty p(t)(1-p(t))g^2(t)v(t)dt\right)} \ . \quad (20.3)$$

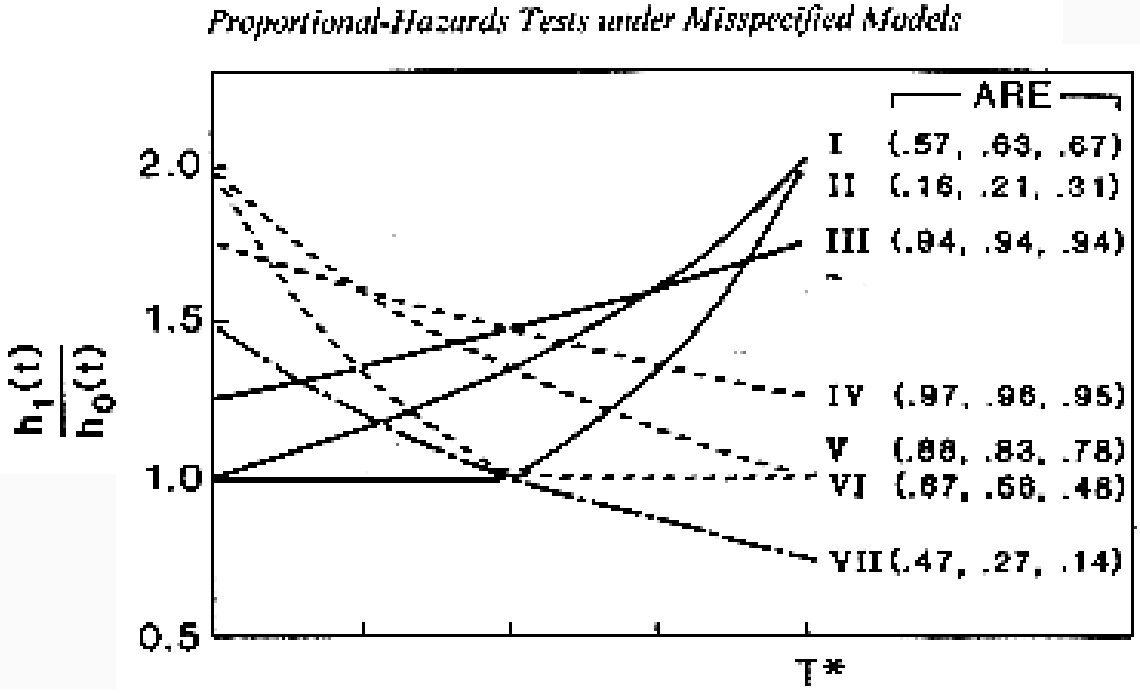
Note that the 2 integrands in the denominator are non-negative for all t, whereas the integrand in the numerator has the same sign as $g(t) \propto \ln(\lambda_1(t)/\lambda_0(t))$. Thus, the logrank test could be inefficient if $\lambda_1(t)/\lambda_0(t) > 1$ for some t and < 1 for others (crossing hazards). It could also be inefficient in other cases.

Figure 20-2 (Lagakos and Schoenfeld, 1984) gives this ARE for a variety of shapes for $\lambda_0(t)$ and $\lambda_1(t)$. All assume $p = P(Z_i = 1) = 1/2$, $\lambda_0(t) = 1$, and

$$G_j(t) = \begin{cases} 0 & t < T^* \\ 1 & t \geq T^* \end{cases} \qquad j = 0, 1$$

where T^* is chosen to give 10%, 25%, or 50% censoring in group 0.
The 3 AREs after each hazard ratio correspond to 10%, 25%, and 50% censoring.

Figure 20-2



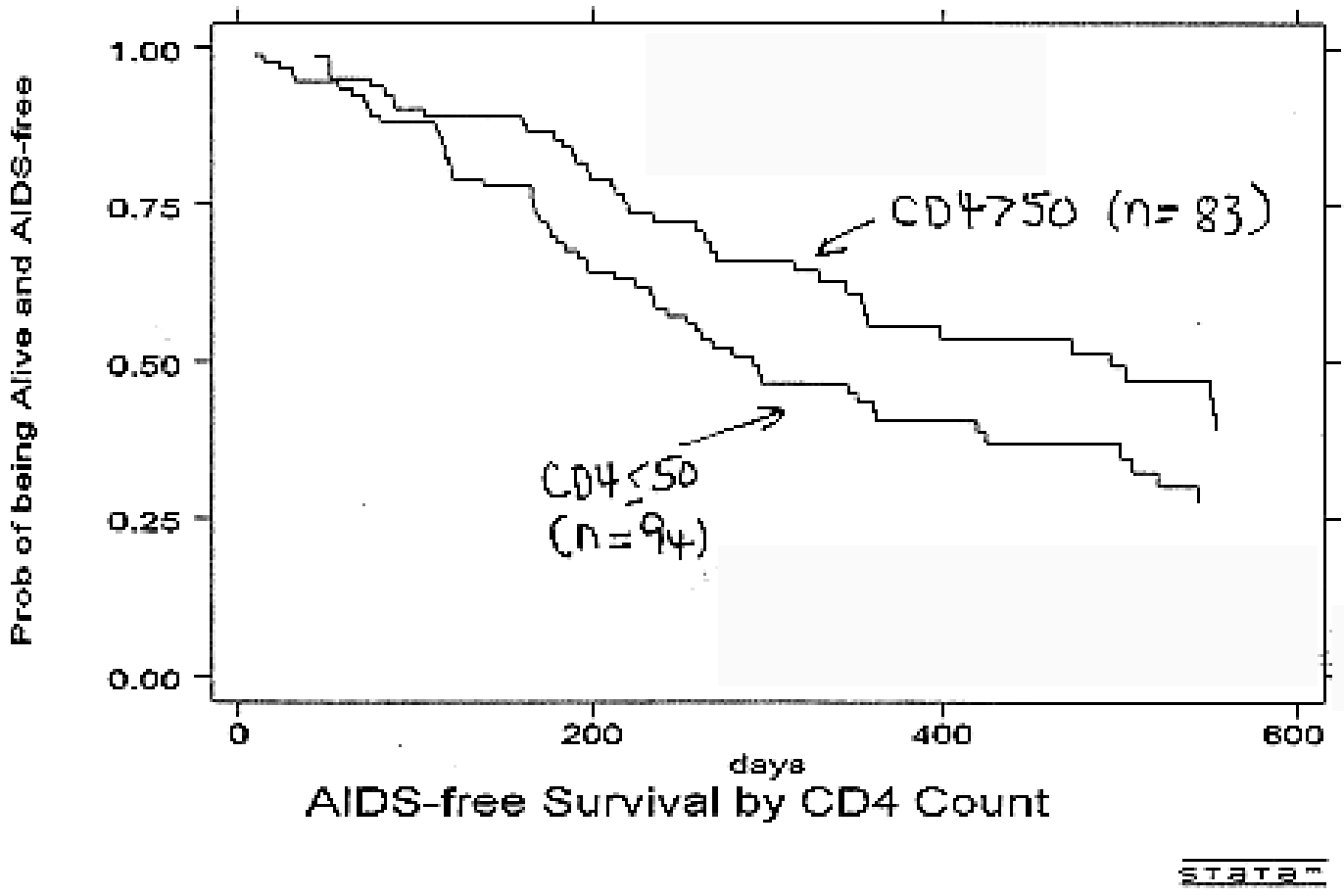
The 3 AREs after each hazard ratio correspond to 10%, 25%, & 50% censoring. Here $p = 1/2$ (equal allocation), $G_0 = G_1$ = type I censoring at T^*.

Note that the logrank tests maintains a pretty high ARE for hazard ratios depicted in III, IV, and V, does less well for I and VI, and does poorly for the crossing hazard functions depicted in VII. Keep in mind that this is the ARE of Z_{LR} relative to the optimal weighted logrank test, which is not really attainable in practice. Thus, unless there is good reason to believe that the hazard ratio departs strongly from a constant, it is reasonable to use the ordinary logrank test and be confident that it is not highly inefficient relative to a weighted logrank test.

Another obvious comparison is the logrank test versus Z_{GW}, the Generalized Wilcoxon test with $w_n(s) = \hat{S}(s-)$. This is discussed in the exercises.

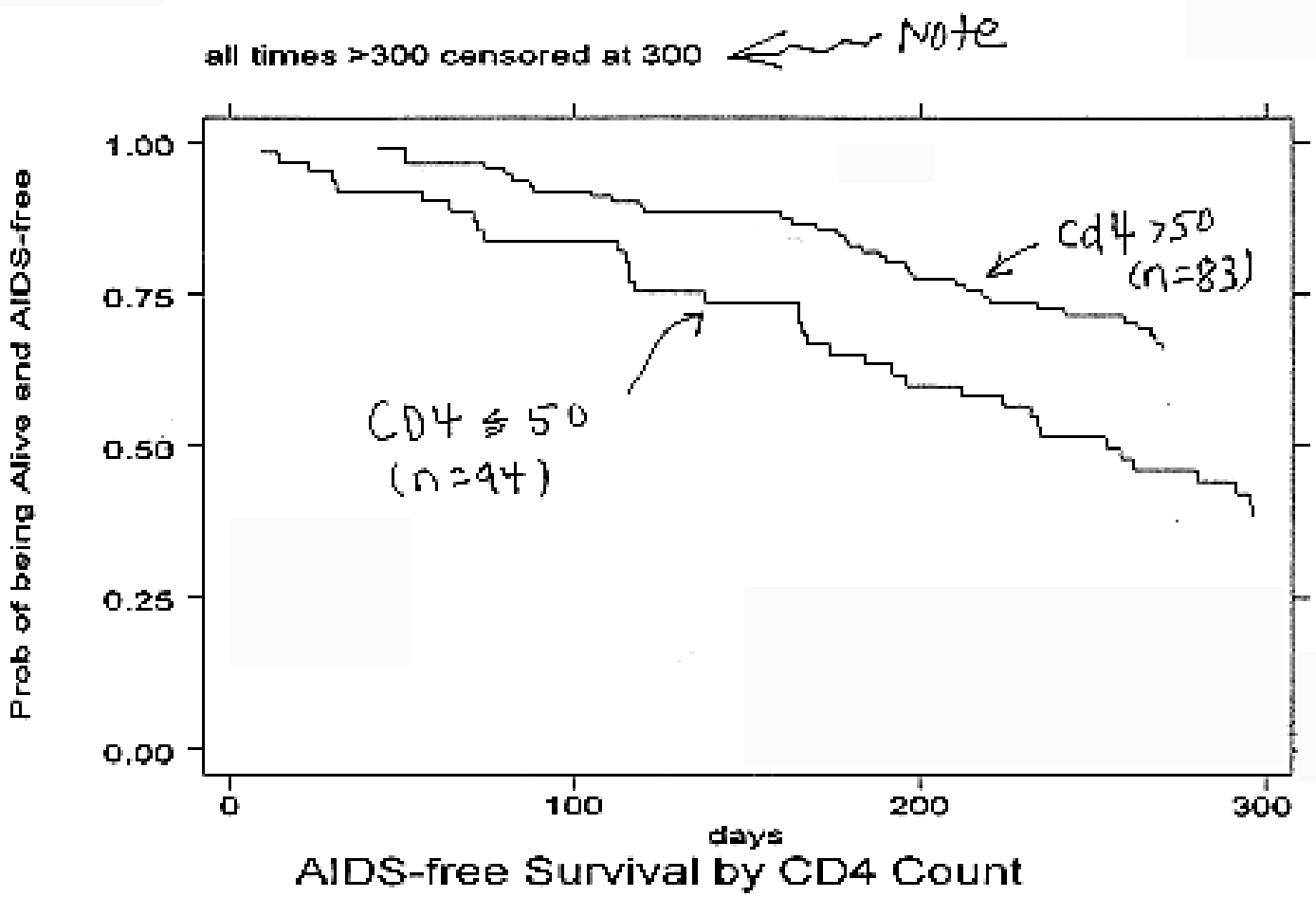
In theory, we know that Z_{GW} is oriented to alternatives to H_0 where the hazard ratio is decreasing with time. In practice, it's not often clear, a priori, when this is likely to happen. Figures 20-3 and 20-4 illustrate 2 examples where Z_{LR} and Z_{GW} give similar results; this is not uncommon. The Exercises shed some light on the reasons for this.

Figure 20-3: from ex3.dta



Logrank Test: p=0.0345
Wilcoxon Test: p=0.0227

Figure 20-4: from ex3.dta



Logrank Test: p=0.0141
Wilcoxon Test: p=0.0176

Sample Size Derivation

Equation (20.1) can also be used to plan the size/duration of a study. For example, suppose we want to have a certain power to detect a specific alternative. Then, for given values of $f_0(t)$, $\ln(\lambda_1(t)/\lambda_0(t))$, G_0 and G_1 (these are determined by the accrual rate and duration of the trial), p, and $w(t)$, we can compute the NCP θ in (20.1), replacing $g(t)$ by $n^{1/2}\ln(\lambda_1(t)/\lambda_0(t))$.

We can then find the actual value of θ needed to give the desired power by assuming that Z_w is $\mathcal{N}(\theta, 1)$ under H_A and then solving for n. For the special case where $p = 1/2$, $G_0(\cdot) = G_1(\cdot)$, a logrank test is used ($w(t) = 1$), and $\lambda_1(t)/\lambda_0(t) = \rho$, (20.1) simplifies to

$$\theta = ln(\rho)\sqrt{nP(\delta = 1)}/2.$$

To have 90% power to detect such a difference, for $\rho > 1$, using a 2-sided Type I error of 0.05, we need $\theta = 1.96 + 1.28 = 3.24$, and thus if $\rho = 1.5$,

$$nP(\delta = 1) = \frac{4\theta^2}{(\ln(\rho))^2} = \frac{4(3.24)^2}{ln(1.5)^2} = 255.$$

That is, since the left-hand size is the expected number of observed failures, one would need a trial with n=255 observed failures to have 90% power to detect a hazard ratio of 1.5 at a Type I level of 0.05. This approximate is very close to what standard sample size packages, such as EAST, give for this setting. Note that the actual number of patients needed would be greater than 255 since not all enrolled patients would be observed to fail.

Alternatively, we could fix n and vary the length of the study (i.e., vary G_0 and G_1) to give the desired numerical value of θ.

These calculations are much simplified when assuming that the log hazard ratio $\lambda_1(t)/\lambda_0(t)$ is constant and when we are using a logrank test ($w(t) = 1$). However, (20.1) can be used for any situation.

Exercises

1. Verify the formula for θ in (20.1). You will need to refer to Unit 19.

2. Suppose that $G_1(\cdot) = G_0(\cdot)$.

   (a) Find an expression for the ARE of the Generalized Wilcoxon Test ($w(s) = \hat{S}(s-)$) to the Logrank test when the 2 groups being compared have proportional hazards.

   (b) Simplify the expression in (a) further by assuming $F_0(\cdot) \sim NE(1)$ and $G_0(\cdot) \sim NE(\varphi)$.

   (c) Evaluate, for $\varphi = .5, 1, 2, 4, 5$, the ARE in (b) as well as $P(\delta_i = 0 \mid H_0)$.

   (d) Is the trend you see in (c) surprising in light of the formula in (a) and the fact that $S(t)$ goes from 1 to 0 as $t \to \infty$?

3. Consider the 2-sample problem with noninformative right censoring and the use of the (ordinary) logrank test to compare the survival distributions of the 2 groups.

   (a) What can you say about the performance of the logrank test compared to a parametric analysis which assumes that the survival distributions in the 2 groups have exponential distributions?

   (b) What can you say about the performance of the logrank test when the true hazard ratio is non-proportional?

4. Use (20.2) to prove that if $G_0(\cdot) = G_1(\cdot)$, ARE=1 for the exponential distribution.

5. Consider a situation in which X denotes treatment group (0 or 1). Suppose that the distribution of censoring does not depend on X, and that the observations are noninformatively right censored. Let p be the probability that $X = 1$. Suppose moreover that the hazard of T, the failure time, is given by

$$\lambda(t|X) = \lambda_0(t)e^{\alpha X}.$$

Find an expression for the Non-Centrality Parameter (NCP) of the log-rank test that compares treatment groups.

References

Fleming TR & Harrington DP (1991). Counting Processes and Survival Data, Wiley, New York

Lagakos SW & Schoenfeld DA (1984). *Biometrics*, p. 1037-

Schoenfeld DA. (1981), *Biometrika* vol. 68, no. 1, p. 316-319