Chapter 4

Non-parametric estimation for basic survival models

4.1 Background

4.1.1 Introduction to non-parametric estimation

Hazard rates are natural objects for modelling, but not for estimation. The problem is the same as for density estimation: An observation occurs at an single dimensionless point. The best summary of the actual observations would be to say that the density is infinite at the points of observation, and zero elsewhere. Various assumptions of smoothness of the density or hazard rate will lead to optimal smoothing procedures for hazard-rate estimation. (An alternative approach, called *isotonic estimation*, is to impose assumptions like increasing hazard rate, which indirectly force a certain degree of smoothness.) We will not have much to say about hazard-rate (or density) estimation in this course.

Instead, we will be concerned with estimating cumulative hazard rates, which are equivalent to estimating survival functions (or cdfs), since $S(t) = e^{-\Lambda(t)}$. Whereas hazard rates can be any nonnegative function, cumulative hazard rates are by definition increasing functions, which is a much more manageable class of functions to search.

4.1.2 The multiplicative intensity model

Our basic model is a counting process that is a sum of n individual counting processes

$$N(t) = \sum_{i=1}^{n} N_i(t), \qquad \text{where } N_i(t) \text{ has intensity } \lambda_i(t) = \alpha(t)Y_i(t),$$

and $Y_i(t)$ is a random process that is 0 or 1, depending on whether individual i is "at risk" up to time t. The intensity of $N(t)$ is then $\lambda(t) = \alpha(t)Y(t)$, where

$$Y(t) = \sum_{i=1}^{n} Y_i(t) = \#\text{ individuals at risk at time } t.$$

Here $\alpha(t)$ is an arbitrary (deterministic) positive function with $\int_0^\infty \alpha(t)dt = \infty$.

We assume that $Y_i(t)$ is predictable; that is, $Y_i(t)$ is known infinitesimally before time t. In particular, there is no hidden frailty or other unknown information for individual i, and $Y_i(t)$ doesn't depend on the jump that happens or doesn't happen at time t. Our goal is to estimate $A(t) = \int_0^t \alpha(s)ds$.

In the survival setting — where each individual can have at most one event T_i — $Y_i(t) = 0$ for $t > T_i$, and $\alpha(t)$ represents the instantaneous probability of the event occurring for an at-risk individual.

This framework allows for both left truncation and right censoring. Left truncation means that events that happen before a certain (possibly random) time t_i exclude the individual from the study. This is represented in the model by $Y_i(t)$ that starts at 0, and then jumps to 1 at time τ_i if and only if $\tau_i < T_i$. This means that a truncated individual is represented by an intensity that is always 0, which is equivalent (from the point of view of estimating the intensity α) to not being included in the study.

Right censoring means that observations after the (possibly random) censoring time C_i are not observed. This is represented by a Y_i function that starts at 1, and then drops to 0 at $C_i \wedge T_i$.

Note the asymmetry: In censoring, we always observe either the event time or the censoring time, but never both.[1] In truncation, we observe either

[1] In theory. In practice, the censoring time may be difficult to observe. If a survival time is censored because a subject has moved away — or died of an unrelated cause — this may not be known immediately, or ever. If there is, say, a five-year follow-up, at which point it becomes known that a certain subject moved away and left no forwarding address, there may be no way to determine at precisely what point he or she became unobservable.

both times — truncation time followed by event time for left truncation — or neither, which is why the latter are equivalent to not being included in the study at all.[2] Of course, both may be active: An individual whose left-truncation time precedes the right-censoring time has a Y_i that jumps from 0 up to 1 at time τ_i, and then down to 0 at time $C_i \wedge T_i$, only if $\tau_i < C_i \wedge T_i$. If $\tau_i \geq C_i \wedge T_i$ then $Y_i(t)$ is always 0.

These are not the only possibilities for Y_i. It is possible for individual i to be at risk only at certain times (for instance, if the event time is the completion of a certain task, and the subject takes breaks, during which $Y_i = 0$). It may be that the maximum number of events by the individual is two or more (perhaps random), in which case $Y_i(t)$ remains at 1 until that number of events has been completed. $Y_i(t)$ may depend in a complicated way on the other $N_j(t)$. For instance, in a matched case-control study the subjects may be paired up, with observations being made until the first of the two has an event. In this case, we would define $Y_i(t) = \mathbf{1}_{\{t \leq T_j \wedge T_i\}}$, where j is the partner of i.

4.1.3 Martingale analysis of the multiplicative intensity model

Suppose there are n individuals at risk, and let $M^{(n)}(t) = n^{-1/2}\big(N(t) - \Lambda(t)\big)$, where $\Lambda(t) = \int_0^t Y(s)\alpha(s)ds$ is the random cumulative intensity. Following section 3.5.4 we know that $M^{(n)}$ is a martingale with predictable variation $\langle M^{(n)} \rangle(t) = \Lambda(t)/n$.

There may be some complicated rule for determining whether an individual i is at risk at time s, but whatever it is, there is some number $v(s) = \mathbb{P}\{Y_i(s) = 1\}\alpha(s)$. We apply Theorem 3.6 with this choice of v and $H^{(n)} = n^{-1/2}$. Clearly (3.15) is satisfied, and

$$\langle M^{(n)} \rangle(t) = n^{-1} \int_0^t \sum_{i=1}^n Y_i(s)\alpha(s)ds \xrightarrow{n\to\infty} \mathbb{E}[Y_i(s)\alpha(s)] = \int_0^t v(s)ds =: V(t)$$

by the weak law of large numbers. We may conclude that $(N(t) - \Lambda(t))/\sqrt{nV(t)}$ converges to standard normal as $n \to \infty$. Note that $N(t)$ is known, and what we want to estimate is $A(t)$, which is related to $\Lambda(t)$ in a computable way.

Recording five years as the censoring time will clearly overestimate the time at risk, and so underestimate the hazard rate.

[2] Again, in theory. In practice we may not know precisely when an individual in our study first was at risk, from the perspective of the study. Fortunately, the standard methods do not require that we know anything about the individuals who were truncated out of the study.

This means that estimates of $\Lambda(t)$ may be expected to be approximately normal for large n.

This is not a practical result, though, since there are too many unknown quantities mixed up together. We improve on this approach in section 4.2.

4.2 The Nelson–Aalen estimator

4.2.1 Distinct event times: Informal derivation

Consider the case of right-censored (but not truncated) survival data that are precisely observed, so that there is no possibility of an exact tie. Let $t_1 < t_2 < \cdots < t_m$ be the (ordered) times at which an event is observed. If there are n individuals under observation, then of course $m \leq n$.

Split up the time period under observation into equal subintervals of width $\epsilon > 0$, each small enough that the probability of two events in any subinterval is negligible. The cumulative hazard may be thought of as approximately

$$A(t) \approx \epsilon \sum_{k=0}^{\lfloor t/\epsilon \rfloor} \alpha(k\epsilon),$$

and $\epsilon\alpha(k\epsilon)Y(k\epsilon)$ is (up to errors of order ϵ^2) the probability of an event occurring in the time interval $[k\epsilon, (k+1)\epsilon]$, conditioned on the past up to time t_k. That is, conditioned on the past $\mathbf{1}_{\{N((k+1)\epsilon)>N(k\epsilon)\}}$ is a Bernoulli random variable with probability $\epsilon\alpha(k\epsilon)Y(k\epsilon)$, for which the natural estimator is simply the random variable itself $\mathbf{1}_{\{N((k+1)\epsilon)>N(k\epsilon)\}}$. Thus, the estimator of $\alpha(k\epsilon)$ is

$$\hat{\alpha}(k\epsilon) = \frac{\mathbf{1}_{\{N((k+1)\epsilon)>N(k\epsilon)\}}}{\epsilon Y(k\epsilon)},$$

which is nonzero just for those intervals at which some $t_i \in [k\epsilon, (k+1)\epsilon)$, at which point it has the value $1/\epsilon Y(t_i)$ (up to an error of order ϵ). Thus

$$\hat{A}(t) = \epsilon \sum_{k=0}^{\lfloor t/\epsilon \rfloor} \hat{\alpha}(k\epsilon) = \sum_{i:t_i \leq t} \frac{1}{Y(t_i)}, \tag{4.1}$$

which is the Nelson–Aalen estimator.

4.2.2 Distinct event times: Formal derivation of the Nelson–Aalen estimator

Since $N(t)$ is a counting process with intensity $\alpha(t)Y(t)$, we may write (informally)

$$dN(t) = \alpha(t)Y(t)dt + dM(t),$$

where $M(t)$ is a martingale jump process. As long as $Y(t)$ stays nonzero[3], we may write

$$\int_0^t \frac{dN(s)}{Y(s)} = \int_0^t \alpha(s)ds + \int_0^t \frac{dM(s)}{Y(s)}.$$

Recall that the integral of a counting process is the same as summing the integrand over the jump points. Thus the left-hand-side is just the Nelson–Aalen estimator $\hat{A}(t)$. The first term on the right-hand side is $A(t)$. Thus

$$\hat{A}(t) - A(t) = \int_0^t \frac{dM(s)}{Y(s)}.$$

Thus $\hat{A}(t) - A(t)$ is a martingale, implying in particular that its expectation is 0 for any t. This means that $\hat{A}(t)$ is an unbiased estimator for $A(t)$ for each t.

4.2.3 Pointwise confidence intervals

From (3.11) we see immediately that

$$\left[\hat{A} - A\right](t) = \int_0^t \frac{1}{Y(s)^2} dN(s).$$

Using formula 3.9,

$$\hat{\sigma}^2(t) := \int_0^t \frac{1}{Y(s)^2} dN(s) = \sum_{t_i \le t} \frac{1}{Y(t_i)^2}, \tag{4.2}$$

is an unbiased estimator for the variance of $\hat{A}(t)$ for each t. Note that the variance will be a sum of $O(n)$ random terms, each of which is on the order of $1/n^2$, so the variance will be like constant $/n$ for large n. In particular, the variance goes to 0, and the estimator is consistent.

[3] We may deal with the problem of $Y(t) = 0$ by integrating only over s such that $Y(s) > 0$; this is done formally in [ABG08], section 3.1.5.

Now we apply Theorem 3.6 to show that $\hat{A}(t)$ is approximately normal. Define

$$\widetilde{M}^{(n)}(t) := \sqrt{n}\big(\hat{A}(t) - A(t)\big) = \int H^{(n)}(t)dM^{(n)}(t),$$

where $H^{(n)}(t) = \sqrt{n}/Y^{(n)}(t)$ and $M^{(n)}$ is a counting-process martingale with intensity $\lambda^{(n)}(t) = Y^{(n)}(t)\alpha(t)$ (corresponding to n individuals). Let $y(s) := \mathbb{P}\{Y_i(s) = 1\}$. By the Weak Law of Large Numbers

$$\frac{Y^{(n)}(t)}{n} \xrightarrow{n\to\infty}_P y(t) \text{ for each } t,$$

which implies that $H^{(n)}(t) \xrightarrow{n\to\infty}_P 0$ and

$$H^{(n)}(t)^2\lambda^{(n)}(t) = \frac{\alpha(t)}{Y^{(n)}(t)/n} \xrightarrow{n\to\infty}_P \frac{\alpha(t)}{y(t)}.$$

So both conditions are satisfied, and we may conclude that $\hat{A}(t)$ is approximately normal with mean $A(t)$ and variance $v(t) = \int_0^t \alpha(s)/y(s)ds$. Of course, we cannot compute this, because we don't know α or y, but we have the estimator (4.2) for the variance.

Thus, we may write an approximate $(1-\alpha)100\%$ confidence interval for $A(t)$ as

$$\sum_{t_i\le t}\frac{1}{Y(t_i)} \pm z_{1-\alpha/2}\Big(\sum_{t_i\le t}\frac{1}{Y(t_i)^2}\Big)^{1/2}. \tag{4.3}$$

4.2.4 Simulated data set

Suppose that we have 10 observations in the data set with failure times as follows:

$$21, 47, 52, 58+, 71, 72+, 125, 143+, 143+, 143+ \tag{4.4}$$

Here + indicates a censored observation. Then we can calculate the Nelson–Aalen estimator for $S(t)$ at all time points. It is obviously unsafe to extrapolate much beyond the last time point, 143, even with a large data set.

4.2.5 Breaking ties

We describe here only the survival setting, where each individual has a maximum of one event, so that $Y_i(t) = 0$ for $t \ge T_i$.

Table 4.1: Computations of survival estimates for simulated data set (4.4)

t_i	$Y(t_i)$	$1/Y(t_i)$	$\hat{A}(t_i)$	$\widetilde{S}(t_i)$
21	10	0.100	0.100	0.90
47	9	0.111	0.211	0.81
52	8	0.125	0.336	0.71
71	6	0.167	0.503	0.60
125	4	0.25	0.753	0.47

We are assuming here that events have continuous distributions, though rounding may lead to nominal ties. This has important implications for how we deal with ties that appear in our data.

By convention we always assume that censoring follows an event when they are nominally simultaneous. The argument is that $Y(T_i)$ is the number of individuals at risk *just before* time T_i.

When multiple events are simultaneous, we treat them as though they were in fact distinct, even if the distinction is unknown. Suppose d_i is the number of events reported at time t_i. If $Y(t_i)$ is the number at risk just before t_i, then the first death lowers this to $Y(t_i) - 1$, then $Y(t_i) - 2$, and so on, until it reaches $Y(t_i) - d_i$, which is the same as $Y(t_i+) = \lim_{\delta \downarrow 0} Y(t_i + \delta)$. This changes the estimator for $A(t)$ to

$$\hat{A}(t) = \sum_{t_i \le t} \sum_{k=0}^{d_i - 1} \frac{1}{Y(t_i) - k}. \tag{4.5}$$

The corresponding variance is then

$$\mathrm{Var}(\hat{A}(t)) = \sum_{t_i \le t} \sum_{k=0}^{d_i - 1} \frac{1}{(Y(t_i) - k)^2}. \tag{4.6}$$

Obviously this is somewhat crude. There is no mathematical theory behind it. In order to make a precise theory out of this we need a model for how the ties are arising. In the extreme case, we might view the ties as being "real ties", leading to a slightly different estimator — described in section 3.1.3 of [ABG08] — but one that will in practice differ only very slightly.

Simulated example with ties

Suppose that we now have 10 observations in the data set with failure times as follows:

$$21, 47, 47, 58+, 71, 71+, 125, 143+, 143+, 143+ \tag{4.7}$$

The tie between the event at 71 and the censoring at 71 is resolved by assuming the event happens first, leaving the estimators unchanged. The effect of the tie at 47 is to reduce the number of jumps, but our resolution of the tie leaves the estimator unchanged outside of the interval where the observations have changed. The resulting Nelson–Aalen estimator is computed in Table 4.2, and plotted in Figure 4.1.

Table 4.2: Computations of survival estimates for simulated data set (4.7)

t_i	$Y(t_i)$	hazard increment	$\hat{A}(t_i)$	$\widetilde{S}(t_i) = e^{-\hat{A}(t_i)}$
21	10	0.100	0.100	0.90
47	9	$0.111 + 0.125$	0.336	0.71
71	6	0.167	0.503	0.60
125	4	0.25	0.753	0.47

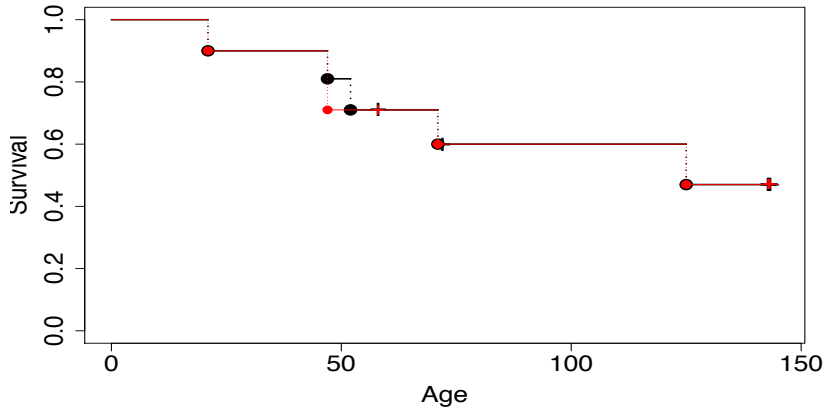


Figure 4.1: Plot of Nelson–Aalen survival estimates from Table 4.1 (black) and Table 4.2 (red).

4.2.6 More simulated data

Consider a large number n of individuals under observation with hazard rate $\alpha(t) = t$ at time t. Suppose they are randomly right-censored with constant rate 1, starting from time $t = 1$. The probability of an individual still being at risk at time t is

$$y(t) = \mathbb{P}\{T > t\}\mathbb{P}\{C > t\} = \begin{cases} e^{-t^2/2} & \text{if } t \le 1, \\ e^{-t^2/2-t+1} & \text{if } t > 1. \end{cases}$$

Thus, by the calculations of section 4.2.3, the asymptotic variance at time t is

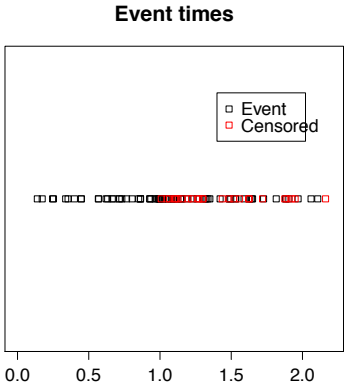
$$v(t) = \int_0^t \frac{\alpha(s)}{y(s)} ds = \int_0^t s e^{s^2/2+(s-1)_+} ds.$$

In Figure 4.2(b) we show a single realisation of this simulation with $n = 100$. The black curve is the Nelson–Aalen estimate for the cumulative hazard, the red curve is the true cumulative hazard $H(t) = t^2/2$. The dashed curves show pointwise upper and lower 95% confidence limits, computed from (4.2). This particular realisation had 75 observed event times, and 25 censored times, shown in Figure 4.2(a). In Figure 4.2(c) we show 20 different additional realisations of the Nelson–Aalen estimator, based on 20 new simulations of the data, together with the original confidence limits for the first simulation. We see that the estimates mostly stay within the confidence limits, but occasionally go outside.

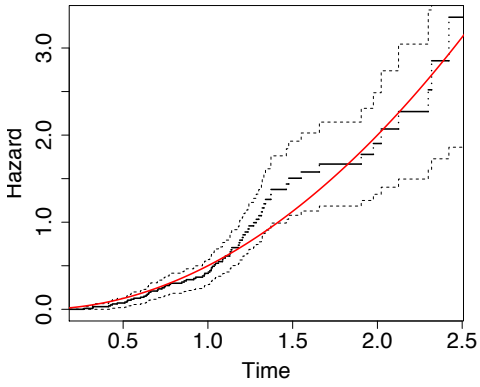
In Figure 4.4 we examine our theoretical variances. The red curve is the known cumulative hazard; the green and blue dashed curves represent $\pm\sqrt{v(t)}/10$ and $\pm\sqrt{v(t)}/5$ — so, $\pm$ 1 or 2 SDs. At times $t = i/10$ for $i = 1, \ldots, 20$ we plot Nelson–Aalen estimators from 100 independent samples. We observe the number of estimates that fall outside these ranges as approximately what we would have expected. In Figure **??** we look at just 20 simulated populations, but we connect up the estimators for different times corresponding to the same simulation, showing how many of the estimators leave the confidence intervals at some time.

In Figures 4.5(a) and 4.5(b) we examine our data-derived estimates for the variance. We simulate the survival process 100 times, and compute the estimated variance $\hat{\sigma}^2(t)$ at times $t = i/10$, $i = 1, 2, \ldots, 20$. These are plotted, together with the known variance $v(t)/n$. We see that the red curve does lie approximately in the middle of each column of estimated variances, consistent with the fact that $\hat{\sigma}^2(t)$ is an unbiased estimator for the variance. On the other hand, the errors are not inconsequential. A more careful analysis would use a Student-like distribution instead of normal for the confidence intervals, to allow for the uncertainty in the variance. It is not entirely straightforward, though, and we will ignore this complication.

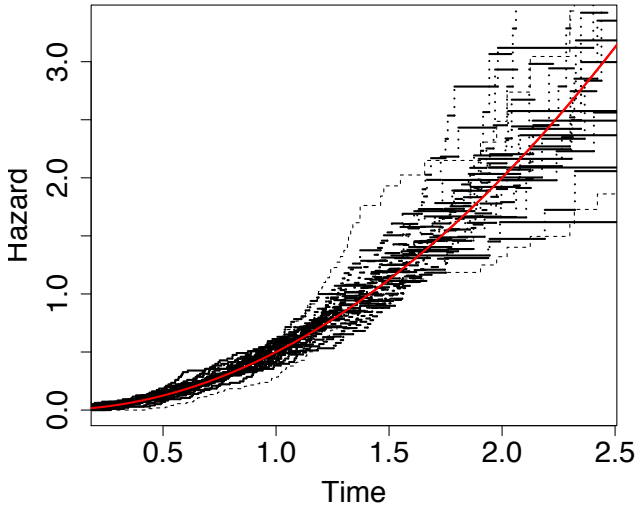
Figures 4.6(a) through 4.6(d) show the distribution of 1000 cumulative hazard estimates at $t = 0.5$ and $t = 1.5$ respectively, as histograms and as normal Q–Q plots. We see that the distributions are approximately normal, but are somewhat right-skewed.



(a) Event times



(b) Single NA estimator with 95% CI



(c) 20 different simulated Nelson–Aalen estimators

Figure 4.2: Nelson–Aalen estimator for 100 simulated individuals with hazard rate t and constant censoring rate 1 after time 1.

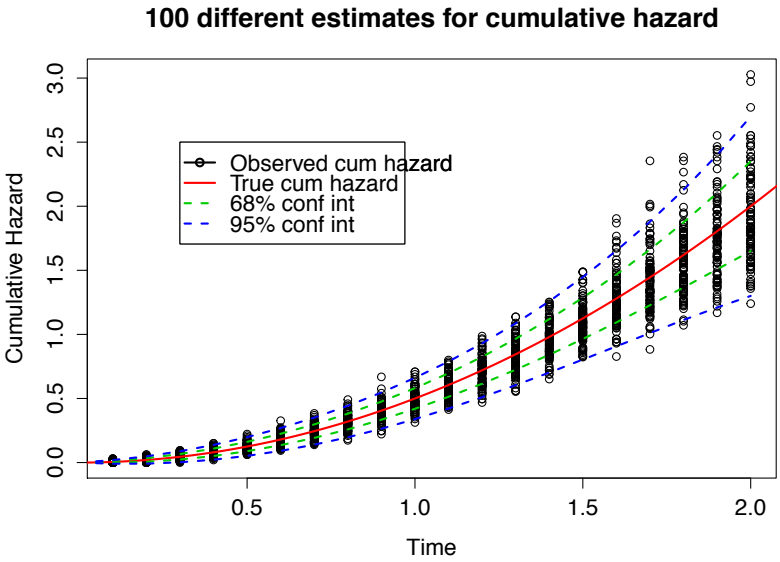


Figure 4.3: 100 Nelson–Aalen estimators for 100 each individuals at specific times, together with theoretical confidence intervals.

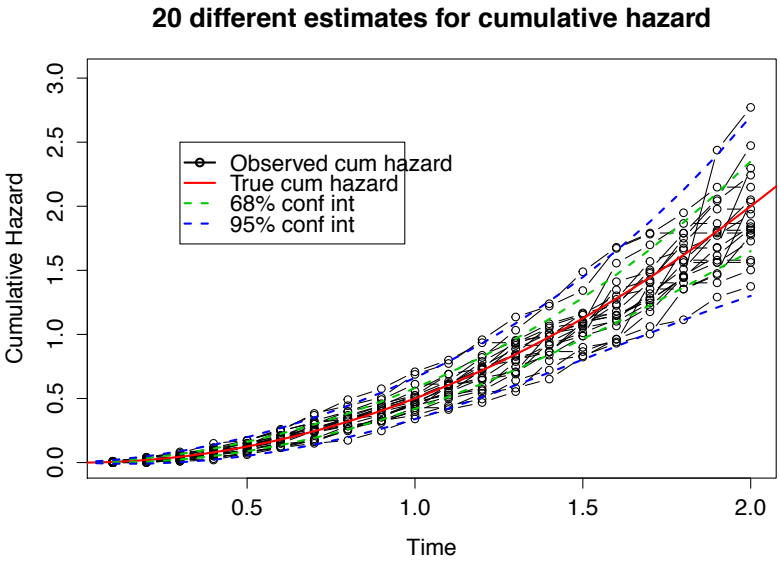
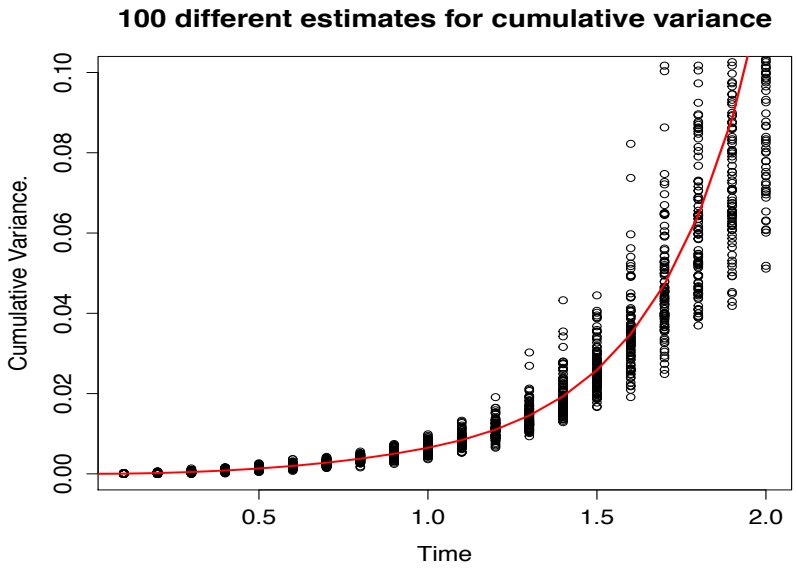
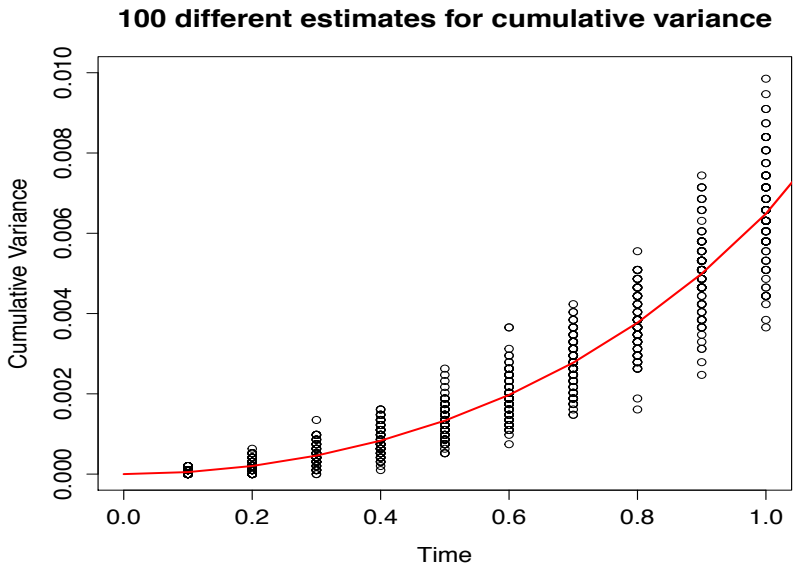


Figure 4.4: 20 Nelson–Aalen estimators for 100 each individuals at specific times, joined up by simulation, together with theoretical confidence intervals.

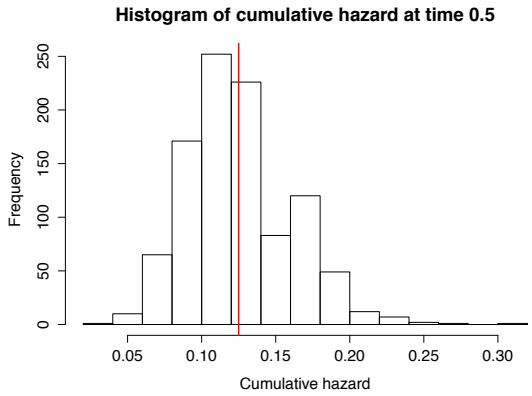


(a) Full time interval

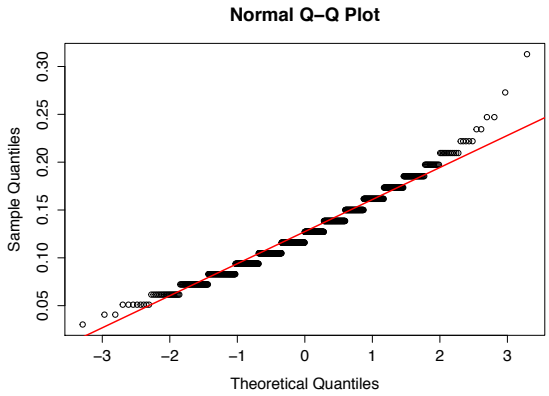


(b) Blowup of low-variance interval

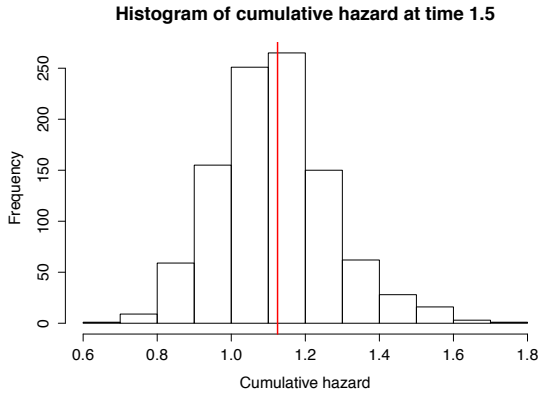
Figure 4.5: 100 variance estimates for different time-points. Red curve shows the true variance.



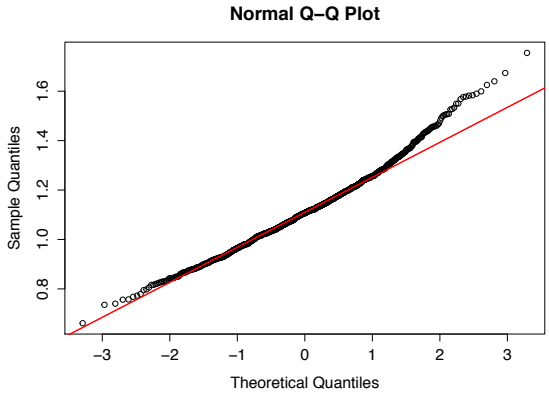
(a) Histogram of time $t = 0.5$ variance estimates



(b) Normal Q-Q plot of time $t = 0.5$ variance estimates



(c) Histogram of time $t = 1.5$ variance estimates



(d) Normal Q-Q plot of time $t = 1.5$ variance estimates

Figure 4.6: Distribution of cumulative hazard estimates at times $t = 0.5$ and $t = 1.5$, based on 1000 simulated populations. Red lines show the true variance.

4.3 The nobody-left problem

In most survival schemes that we consider, there is always at least a small probability that $Y(t)$ will be 0. As discussed at various points in chapter 3 of [ABG08], the solution is to understand that we are "really" estimating not $A(t)$, but rather a quantity

$$A^*(t) := \int_0^t \mathbf{1}_{\{Y(u)>0\}} dA(u). \tag{4.8}$$

With this correction, the mathematics we have done so far becomes rigorously correct. Of course, this adds some confusion because the thing we are estimating is not actually a deterministic quantity, but is itself a random variable. What you should keep in mind is:

(i). $\mathbb{E}[A^*(t) - \hat{A}(t)] = 0$. In that sense, $\hat{A}$ is an "unbiased estimator" of A^*.

(ii). $A^*(t) \leq A(t)$, and it will eventually be strictly less. Thus $\hat{A}$ is, formally, a biased estimator for A.

(iii). $A^*(t)$ and $A(t)$ are genuinely the same, as long as the population hasn't run out.

(iv). But we do often run survival experiments until there is no one left. After that point, we see the flat hazard rate of $\hat{A}$, corresponding to the flat hazard rate of A^*.

The crucial point is the last one: What happens to the Nelson–Aalen estimator at the end is not reflective of any truth about the true hazard rate.

Similarly, the Kaplan–Meier estimator is not an estimator for $S(t)$, but for

$$S^*(t) := \int_0^t \mathbf{1}_{\{Y(u)>0\}} dS(u). \tag{4.9}$$

As we will see, it is slightly more technically important to distinguish between $S(t)$ and $S^*(t)$ when analysing the Kaplan–Meier estimator.

4.4 The Kaplan–Meier estimator

4.4.1 Deriving the Kaplan–Meier estimator

The Nelson–Aalen estimator arises naturally from our mathematical framework, but it is not the most commonly used nonparametric survival estimator.

We consider first the case when the event times are all distinct and possibly right-censored, and recapitulate the informal derivation of the Nelson–Aalen estimator.

Let $t_1 < t_2 < \cdots < t_m$ be the (ordered) times at which an event is observed. If there are n individuals under observation, then of course $m \leq n$.

Split up the time period under observation into equal subintervals of width $\epsilon > 0$, each small enough that the probability of two events in any subinterval is negligible. Instead of estimating the increments to the hazard, we think of estimating the increments to the survival function $S(t) = e^{-A(t)} = \mathbb{P}\{t_i > t\}$. The probability that an individual who has survived to time $(k-1)\epsilon$ has its event before $k\epsilon$ is $1 - S(k\epsilon)/S((k-1)\epsilon)$. If no event occurs in the interval $[(k-1)\epsilon, k\epsilon)$ (though possibly a censoring time), it is natural to estimate $1 - S(k\epsilon)/S((k-1)\epsilon) \approx 0$; if there is a single event t_i — out of $Y((k-1)\epsilon) \approx Y(t_i)$ individuals at risk — it is natural to estimate $1 - S(k\epsilon)/S((k-1)\epsilon) \approx 1/Y(t_i)$. This leads us to the Kaplan–Meier estimator

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{1}{Y(t_i)}\right). \tag{4.10}$$

The fraction $1 - S(t)/S(t-)$ is sometimes referred to as the *discrete hazard* at time t. It is the probability of an individual alive up to time t surviving to time t, where t is a time of discontinuity in the survival function S. Thus, the Kaplan–Meier estimator is the cumulative product up to time t of the empirically estimated discrete hazards up to time t.

If there are ties, then we follow the same strategy as in section 4.2.5 for the Nelson–Aalen estimator. We would then be multiplying together terms of the form

$$\begin{aligned}\prod_{k=0}^{d_i-1} \left(1 - \frac{1}{Y(t_i) - k}\right) &= \left(\frac{Y(t_i) - 1}{Y(t_i)}\right)\left(\frac{Y(t_i) - 2}{Y(t_i) - 1}\right) \cdots \left(\frac{Y(t_i) - d_i}{Y(t_i) - d_i + 1}\right) \\ &= \left(1 - \frac{d_i}{Y(t_i)}\right).\end{aligned}$$

So we have the Kaplan–Meier estimator with ties as

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y(t_i)}\right). \tag{4.11}$$

4.4.2 The relation between Nelson–Aalen and Kaplan–Meier

> If you have one clock you know what time it is. If you have two clocks you are never sure.

— Wisdom whose origins are lost to human memory

We now have two estimators for survival, each of which has a cogent argument in its favour. Which one is "right"?

In sections 3.2.4 and 3.2.6 of [ABG08] there is an interesting discussion of the relationship between the Nelson–Aalen and Kaplan–Meier estimators. There heavy use is made of the "product-integral" notation, a sort of multiplicative version of the integral, represented by a curvy product symbol that requires advanced computing skills to even print in LaTeX.

It's interesting, and worth looking at (not examinable!), but it depends on a non-obvious redefinition of the cumulative hazard in order to make the Nelson–Aalen estimator precisely the cumulative hazard corresponding to the Kaplan–Meier estimator. In fact, it's not completely clear what we should mean by a cumulative hazard with jumps — what exactly is the "hazard" here that is being accumulated? — but the most natural definition is that the cumulative hazard is $A(t) = -\log S(t)$. [ABG08] defines the cumulative hazard instead as the integral of $dS(t)/S(t-)$, which agrees with the above definition when A is differentiable, but not when it has jumps.

Trying to show that these estimators are, in some sense, the same distracts from the main point, which is simply that there is no exclusive criterion for a "best" estimator. Some of the criteria we need to balance against each other are

- Ease of computing from the data;
- Minimising bias;
- Minimising error — e.g., MSE;
- Consistency — error is asymptotically 0;
- Asymptotic normality, or more generally, ability to estimate the distribution of errors.

Minimising bias (or seeking unbiased estimators) sounds good — who wants more *bias*? — but is simply not a very important criterion. In any case, there is no way to make a cumulative hazard estimator that is also unbiased as an estimator of survival. The Kaplan-Meier estimator is typically unbiased for the survival function The Nelson–Aalen estimator is unbiased for cumulative hazard.

So what is the difference between Nelson–Aalen and Kaplan–Meier? If we accept the definition of cumulative hazard $A(t) = -\log S(t)$, the change

in cumulative hazard at a jump-point T_i of the Kaplan–Meier estimator $\hat{S}$ is $-\log \hat{S}(T_i)/\hat{S}(T_i-)$; the Nelson–Aalen estimator, on the other hand, changes by $1 - \hat{S}(T_i)/\hat{S}(T_i-)$ — the estimated discrete hazard at time T_i. Of course, these will be very similar if $\hat{S}(T_i)/\hat{S}(T_i-)$ is close to 1. The differences will be on the order of $1/Y(T_i)^2$, and cumulatively on the order of the smallest number of individuals ever at risk. We note for the future the relation

$$d\hat{A}(t) = \frac{d\hat{S}(t)}{\hat{S}(t-)} \tag{4.12}$$

To summarise:

(i). The Nelson–Aalen estimator is unbiased for cumulative hazard; the Kaplan–Meier estimator is unbiased for survival.

(ii). The Nelson–Aalen estimator is always larger than $-\log$ of the Kaplan–Meier estimator.

(iii). They are not very different, as long as the number of individuals at risk remains large.

(iv). If the number of individuals at risk isn't large, there's no good reason to prefer one over the other, although the Kaplan–Meier estimator is more natural as an estimator of survival, while the Nelson–Aalen estimator is more natural as an estimator of cumulative hazard.

(v). The Nelson–Aalen estimator is easier to work with mathematically...

(vi). ...but the Kaplan–Meier estimator is substantially better known and more widely used, particularly in medical contexts. It's also the default in the `survival` package in `R`.

We will generally treat the two estimators interchangeably, taking $\widetilde{S}(t) = e^{-\hat{A}(t)}$ as an estimator for survival, and $-\log \hat{S}(t)$ (where $\hat{S}$ is the Kaplan–Meier estimator) as an estimator for cumulative hazard, using whichever seems more convenient.

4.4.3 Duhamel's equation

A useful relation for comparing different survival curves may be found by differentiating the ratio:

$$\begin{aligned} d(S_1/S_2) &= \frac{S_2 dS_1 - S_1 dS_2}{S_2^2} \\ &= \frac{S_1}{S_2}\left(\frac{dS_1}{S_1} - \frac{dS_2}{S_2}\right). \end{aligned}$$

As discussed in section 4.4.2, for continuous survival curves $-dS(t)/S(t-)$ is the simply the hazard; otherwise, it is the discrete hazard, and we could simply *define* this as the increment to the cumulative hazard dA. This is the approach taken by [ABG08, section A.1].

We assume that S_2 is differentiable, and integrate both sides (which is the only way that this equation makes formal sense) to obtain *Duhamel's equation*:

$$\begin{aligned} \frac{S_1(t)}{S_2(t)} &= 1 + \int_0^t \frac{S_1(s-)}{S_2(s)}\left(\frac{dS_1(s)}{S_1(s-)} - \frac{dS_2(s)}{S_2(s-)}\right) \\ &= 1 + \int_0^t \frac{S_1(s-)}{S_2(s)}\left(dA_2(s) - dA_1(s)\right), \end{aligned} \tag{4.13}$$

where $-dS(s)/S(s-)$ is the fraction of those alive at time s who die at time s (when s is a discontinuity of S).

When S_1 and S_2 are both continuous (and differentiable) at s this is just calculus. When S_1 or S_2 has a jump at s there is a calculation to be done, which is left as an exercise.

4.4.4 Confidence intervals for the Kaplan–Meier estimator

We may apply Duhamel's equation to $\hat{S}/S^*$, where $\hat{S}$ is the Kaplan–Meier estimator, S is the true survival function, and S^* is the survival function that is decremented only when $Y(t) > 0$, as discussed in section 4.3. Then

$$\frac{\hat{S}(t)}{S^*(t)} = 1 + \int_0^t \frac{\hat{S}(s-)}{S^*(s)}\left(\frac{dS^*(s)}{S^*(s-)} - \frac{d\hat{S}(s)}{\hat{S}(s-)}\right),$$

We are assuming that our underlying survival function S is continuous, so $dS^*/S^* = A^*$; and we already pointed out in (4.12) that $d\hat{S}(s)/\hat{S}(s-) = d\hat{A}(s)$ is the Nelson–Aalen estimator. This gives us

$$\frac{\hat{S}(t)}{S(t)} - 1 = \int_0^t \frac{\hat{S}(s-)}{S^*(s)} d\left(\hat{A} - A^*\right)(s). \tag{4.14}$$

Thus $\hat{S}(t)/S^*(t) - 1$ is a mean-zero martingale, and $\mathbb{E}[\hat{S}(t)/S^*(t)] = 1$. Note that S^* is itself random, since it depends on the random variable $\mathbf{1}_{\{Y(t)>0\}}$. For t small enough that $Y(t) > 0$ with high probability, $S^*(t) = S(t)$ with high probability, so we may conclude that

$$\mathbb{E}\big[\hat{S}(t)\big] \approx S(t),$$

so that $\hat{S}(t)$ is almost unbiased. As with the Nelson–Aalen estimator, $\hat{S}(t)$ obviously becomes biased (when considered as an estimator for $S(t)$) — it overestimates survival — when t becomes large enough for $Y(t) = 0$.

Estimating the variance is not quite as straightforward as for the Nelson–Aalen estimator. If n is moderately large then we can approximate $S^*(t) \approx S(t)$ and $\hat{S}(s-)/S^*(s) \approx 1$, so

$$\frac{\hat{S}(t)}{S(t)} - 1 \approx \hat{A}(t) - A(t),$$

or

$$\operatorname{Var}(\hat{S}(t)) \approx S(t)^2 \operatorname{Var}(\hat{A}(t)).$$

Applying (4.2), the variance of the Kaplan–Meier estimator may be estimated by using

$$\hat{\tau}^2(t) := \hat{S}(t)^2 \sum_{t_i \le t} \frac{1}{Y(t_i)^2}. \tag{4.15}$$

Similarly, when there are ties, the estimator is

$$\hat{\tau}^2(t) := \hat{S}(t)^2 \sum_{t_i \le t} \sum_{k=0}^{d_i - 1} \frac{1}{(Y(t_i) - k)^2}. \tag{4.16}$$

An alternative commonly used variance estimator is *Greenwood's formula*, which takes the form

$$\tilde{\tau}^2(t) := \hat{S}(t)^2 \sum_{t_i \le t} \frac{1}{Y(t_i)(Y(t_i - 1)} \tag{4.17}$$

when there are no ties. In the case of no censoring, this reduces to $\hat{S}(t)(1 - \hat{S}(t))$. Obviously these two estimators are asymptotically (as $n \to \infty$) identical

When there are ties, the Greenwood formula is

$$\tilde{\tau}^2(t) := \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{Y(t_i)(Y(t_i) - d_i)}, \tag{4.18}$$

where, again, d_i is the number of events recorded at time t_i.

4.5 Computing survival estimators in R

The main package for doing survival analysis in R is `survival`. Once the package is installed on your computer, you include `library(survival)` at the start of your Rcode This works with "survival objects", which are created by the `Surv` command with the following syntax:
`Surv(time, event)` or `Surv(time, time2, event,type)`

4.5.1 Survival objects with only right-censoring

We begin by discussing the first version, which may be applied to right-censored (or uncensored) survival data. The individual times (whether censoring or events) are entered as vectors `time`. The vector `event` (of the same length) has values 0 or 1, depending on whether the time is a censoring time or an event time respectively. These alternatives may also be labelled 1 and 2, or `FALSE` and `TRUE`.

For an example, we turn to our tiny simulated data set

$$21, 47, 47, 58+, 71, 71+, 125, 143+, 143+, 143+$$

into a survival object with

`sim.surv` = `Surv`(`c`$(21, 47, 47, 58, 71, 71, 125, 143, 143, 143)$, `c`$(1, 1, 1, 0, 1, 0, 1, 0, 0, 0)$).

Fitting models is done with the `survfit` command. This is designed for comparing distributions, so we need to put in a some sort of covariate. Just for our example, we make a variable `type=rep(1,10)`, so everything has type 1. Then we can write

```
sim.fit=survfit(sim.surv~type,conf.int=.99)
```

and then `plot(sim.fit)`, or

```
plot(sim.fit,main='Kaplan-Meier for simulated data set',
          xlab='Time',ylab='Survival')
```

to plot the Kaplan–Meier estimator of the survival function, as in Figure 4.7. The dashed lines are the Greenwood estimator of a 99% confidence interval. (The default for `conf.int` is 0.95.)

The Nelson–Aalen estimator is not computed by the `survfit` function, but it organises all the information that you would need to compute it. If you want to see what's inside an R object, you can use the `str` command. The output is shown in Figure 4.8.

We can then compute the Nelson–Aalen estimator with a function such as the one in Figure 4.9. This is plotted together with the Kaplan–Meier

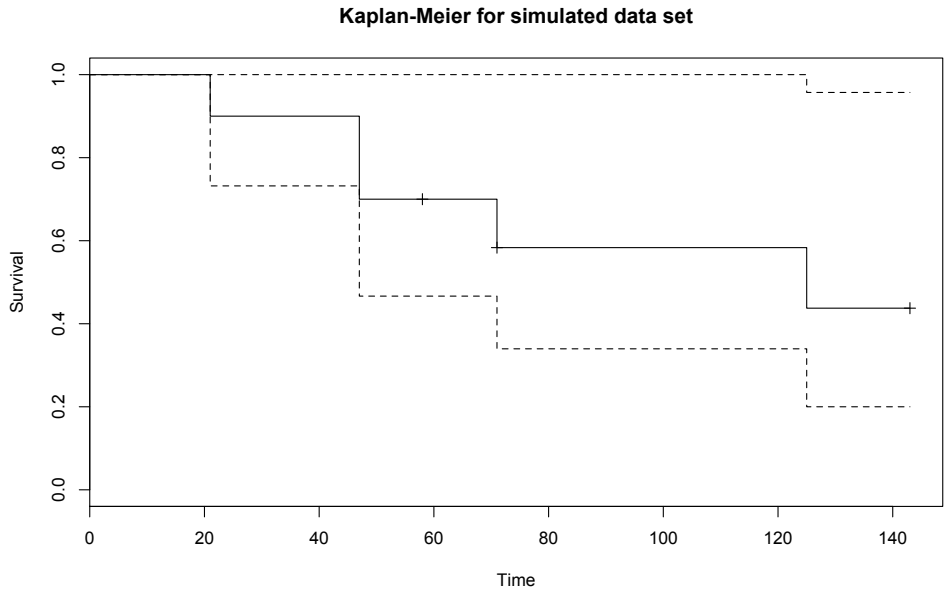


Figure 4.7: Plot of Kaplan–Meier estimates from data in (4.7). Dashed lines are 95% confidence interval from Greenwood's estimate.

estimator in Figure 4.10. As you can see, the two estimators are similar, and the Nelson–Aalen survival is always higher than the KM.

4.5.2 Other survival objects

Left-censored data are represented with
`Surv(time, event,type='left')`.
Here `event` can be 0/1 or 1/2 or `TRUE/FALSE` for alive/dead, i.e., censored/not censored.

Left-truncation is represented with
`Surv(time,time2, event)`.
`event` is as before. The `type` is 'right' by default.

Interval censoring also takes `time` and `time2`, with `type='interval'`. In this case, the `event` can be 0 (right-censored), 1 (event at `time`), 2 (left-censored), or 3 (interval-censored).

```
> str(sim.fit)
List of 13
$ n :  int 10
$ time :  num [1:6] 21 47 58 71 125 143
$ n.risk :  num [1:6] 10 9 7 6 4 3
$ n.event :  num [1:6] 1 2 0 1 1 0
$ n.censor :  num [1:6] 0 0 1 1 0 3
$ surv :  num [1:6] 0.9 0.7 0.7 0.583 0.438...
$ type :  chr ‘‘right’’
$ std.err :  num [1:6] 0.105 0.207 0.207 0.276
0.399 ...
$ upper :  num [1:6] 1 1 1 1 0.957 ...
$ lower :  num [1:6] 0.732 0.467 0.467 0.34 0.2
...
$ conf.type:  chr ‘‘log’’
$ conf.int :  num 0.95
$ call :  language survfit(formula = sim.surv ∼
type, conf.int = 0.95) - attr(*, ‘‘class’’)= chr
‘‘survfit’’
```

Figure 4.8: Example of structure of a `survfit` object.

```
NAest=function(SF){
        times=SF$time[SF$n.event>0]
        events=SF$n.event[SF$n.event>0]
        nrisk=SF$n.risk[SF$n.event>0]
        increment=sapply(seq(length(nrisk)),function(i)
                sum(1/seq(nrisk[i],nrisk[i]-events[i]+1)))
        varianceincrement=sapply(seq(length(nrisk)),function(i)
                        sum(1/seq(nrisk[i],nrisk[i]-events[i]+1)^2))
        hazard=cumsum(increment)
        variance=cumsum(varianceincrement)
        list(time=times,Hazard=hazard,Var=variance)
}
```

Figure 4.9: Function to compute Nelson–Aalen estimator.

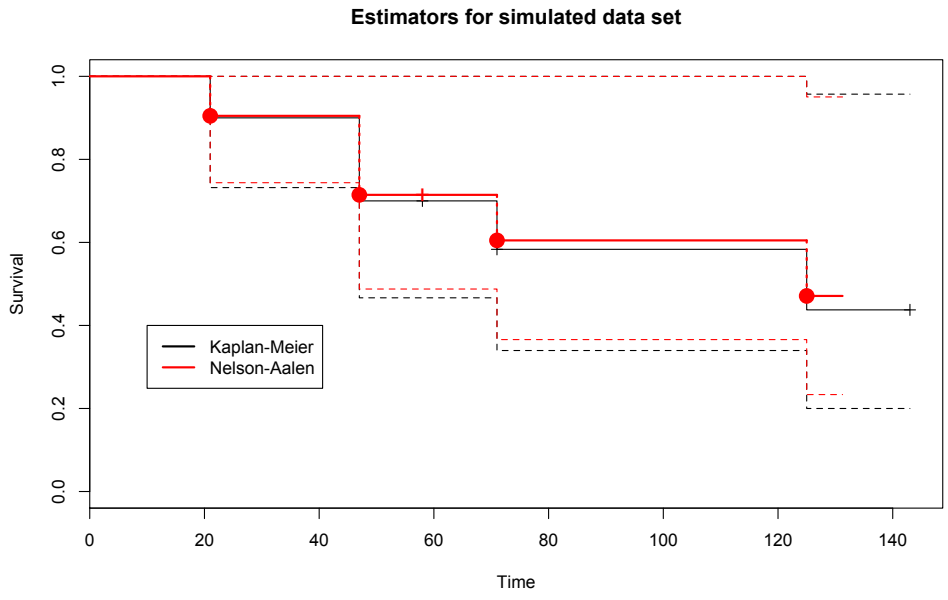


Figure 4.10: Plot of Kaplan–Meier (black) and Nelson–Aalen (red) estimates from data in (4.7). Dashed lines are pointwise 95% confidence intervals.

4.6 Survival to ∞

Let T be a survival time, and define the conditional survival function

$$S_0(t) := \mathbb{P}\big\{T > t \,\big|\, T < \infty\big\};$$

that is, the probability of surviving to time t given that the event eventually does occur. We have

$$S_0(t) = \frac{\mathbb{P}\{\infty > T > t\}}{\mathbb{P}\{\infty > T\}}. \tag{4.19}$$

How can we estimate S_0? Nelson–Aalen estimators will never reach ∞ (which would mean 0 survival); Kaplan–Meier estimators will reach 0 if and only if the last individual at risk actually has an observed event. In either case, there is no mathematical principle for distinguishing between the actual survival to ∞ — that is, the probability that the event never occurs — and simply running out of data. Nonetheless, in many cases there can be good reasons for thinking that there is a time t_∂ such that the event will never happen if it hasn't happened by that time. In that case we may use the fact that $\{T < \infty\} = \{T < t_\partial\}$ to estimate

$$\hat{S}_0 = \frac{S(t) - S(t_\partial)}{1 - S(t_\partial)}. \tag{4.20}$$

Example 4.1: Time to next birth

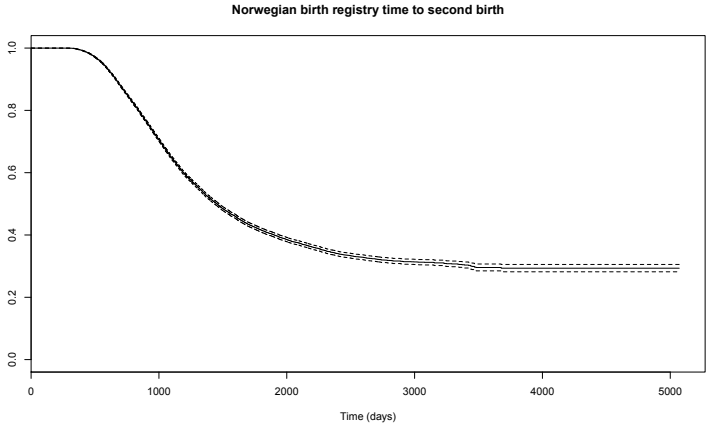
This is an example discussed repeatedly in [ABG08]. It has the advantage of being a large data set, where the asymptotic assumptions may be assumed to hold; it has the corresponding disadvantage that we cannot write down the data or perform calculations by hand.

The data set at `http://folk.uio.no/borgan/abg-2008/data/second_births.txt` lists, for 53,558 women listed in Norway's birth registry, the time (in days) from first to second birth. (Obviously, many women do not have a second birth, and the observations for these women will be treated as censored.)

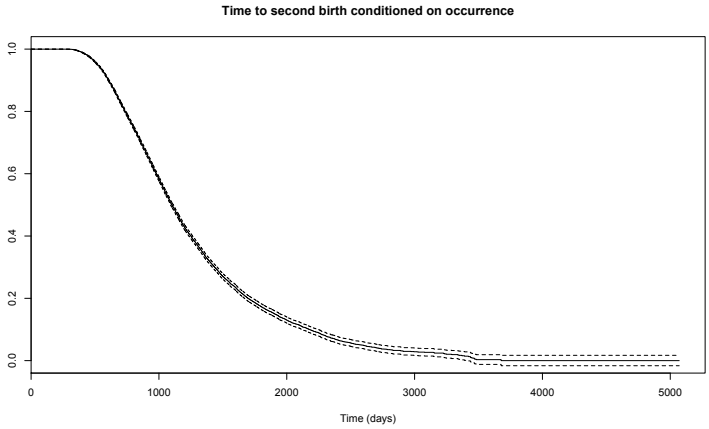
In Figure 4.11(a) we show the Kaplan–Meier estimator computed and automatically plotted by the `survfit` command. Figure 4.11(b) shows a crude estimate for the distribution of time-to-second-birth for those women who actually had a second birth.

We see that the last birth time recorded in the registry was 3677, after which time none of the remaining 131 women had a recorded second birth. Thus, the second curve is simply the same as the first curve, rescaled to go between 1 and 0, rather than between 1 and 0.293 as the original curve does.

The code used to generate the plots is in Figure 4.12. ■



(a) Original Kaplan–Meier curve



(b) Kaplan–Meier curve conditioned on second birth occurring

Figure 4.11: Time (in days) between first and second birth from Norwegian registry data.

```
library('survival')

sb=read.table('second_births.dat',header=TRUE)

attach(sb)

sb.surv=Surv(time,status)

sb.fit1=survfit(sb.surv~rep(1,53558))

plot(sb.fit1,mark.time=FALSE,xlab='Time (days)',
        main='Norwegian birth registry time to second birth')

# Condition on last event
cle=function(SF){
        minsurv=min(SF$surv)
        SF$surv=(SF$surv-minsurv)/(1-minsurv)
        SF$upper=(SF$upper-minsurv)/(1-minsurv)
        SF$lower=(SF$lower-minsurv)/(1-minsurv)
        SF
}

sb.fit2=cle(sb.fit1)

plot(sb.fit2,mark.time=FALSE,xlab='Time (days)',
        main='Time to second birth conditioned on occurrence')
```

Figure 4.12: Code to generate Figure 4.11

.

Chapter 5

Comparing distributions: Excess mortality

5.1 Estimating excess mortality: One-sample setting

This section is taken directly from section 3.2.5 of [ABG08].

A common class of models splits the hazard rate into two pieces:

$$\alpha_i(t) = \gamma(t) + \mu_i(t), \tag{5.1}$$

where $\mu_i(t)$ is a known baseline hazard rate that is associated to individual i, and $\alpha(t)$ is an unknown increment to the hazard that we seek to calculate. For example, we might be measuring mortality, and μ_i the known population mortality for individuals with the same age and gender.

We will discuss in Chapter 7 the problem of testing the validity of such a *relative survival* model, and compare it to the *relative mortality* model that would have $\alpha_i(t) = \mu_i(t)\alpha(t)$. Here, we concern ourselves with how to estimate γ.

We have the intensity for individual i of

$$\lambda_i(t) = \big(\gamma(t) + \mu_i(t)\big)Y_i(t).$$

If we define the *average population mortality*

$$\bar{\mu}(t) = \sum_{i=1}^{n} \mu_i(t)\frac{Y_i(t)}{Y(t)},$$

then the total counting process N has intensity

$$\lambda(t) = \big(\gamma(t) + \bar{\mu}(t)\big)Y(t).$$

If we consider the relative survival function $R(t) = e^{-\Gamma(t)}$, where $\Gamma(t) = \int_0^t \gamma(s)ds$. We then have the martingale representation

$$M(t) := \int_0^t \frac{dN(s)}{Y(s)} - \int_0^t \big(\gamma(s) + \bar{\mu}(s)\big)ds.$$

Thus

$$\hat{\Gamma}(t) := \int_0^t \frac{dN(s)}{Y(s)} - \int_0^t \bar{\mu}(s)ds = \Gamma(t) + M(t)$$

is an unbiased estimator for $\Gamma(t)$. By the same arguments as in section 4.2 we see that $\hat{\Gamma}(t)$ is approximately normal, with variance estimated by $\hat{\sigma}^2(t) = \sum_{t_i \le t} 1/Y(t_i)^2$.

Define the *average survival function*

$$\hat{S}_{ave}(t) := \exp\left\{-\int_0^t \bar{\mu}(s)ds\right\}. \tag{5.2}$$

Observe that this is a random function, since it depends on the random composition of the population. $\hat{S}_{ave}$ is the survival probability that would be observed if the population composition were the one observed, but everyone had the survival rates given by the baseline. Then we may estimate the relative survival function by

$$\hat{R}(t) := \frac{\hat{S}(t)}{\hat{S}_{ave}(t)} \tag{5.3}$$

Example 5.1: Simulated survival data

In Figure 5.1 we plot the mortality rates (from 1990–2) in England and Wales, for male and female separately. Imagine that we had a population of elderly individuals, of various ages, who were observed over 10 years because of their ongoing exposure to an environmental danger. Suppose, in fact, that this pollutant increases mortality rates uniformly by 0.01 per year. How well can we estimate this effect?

We model individual intensities as $\big(\gamma(t)+\mu_{s_i}(a_i+t)\big)Y_i(t)$, where a_i is the initial age of individual i, and s_i is the sex. We simulate a population that starts with 1000 males and 1000 females, with randomly chosen initial ages between 50 and 80 years, and assume censoring at a rate of 0.02 per year.

The results are plotted in Figure 5.2. The red curve shows the estimate $\hat{\Gamma}$, while the black circles show the changing background mortality $\bar{\mu}$. Note that $\bar{\mu}$ grows much more slowly than the mortality rates for individuals, owing to the shift in the population toward younger individuals and women, as the higher-mortality subgroups die off. For example, we start with equal numbers of men and women, but the remaining population after 10 years is close to 60% female. Note, too, that the estimates for Γ are reasonably good.

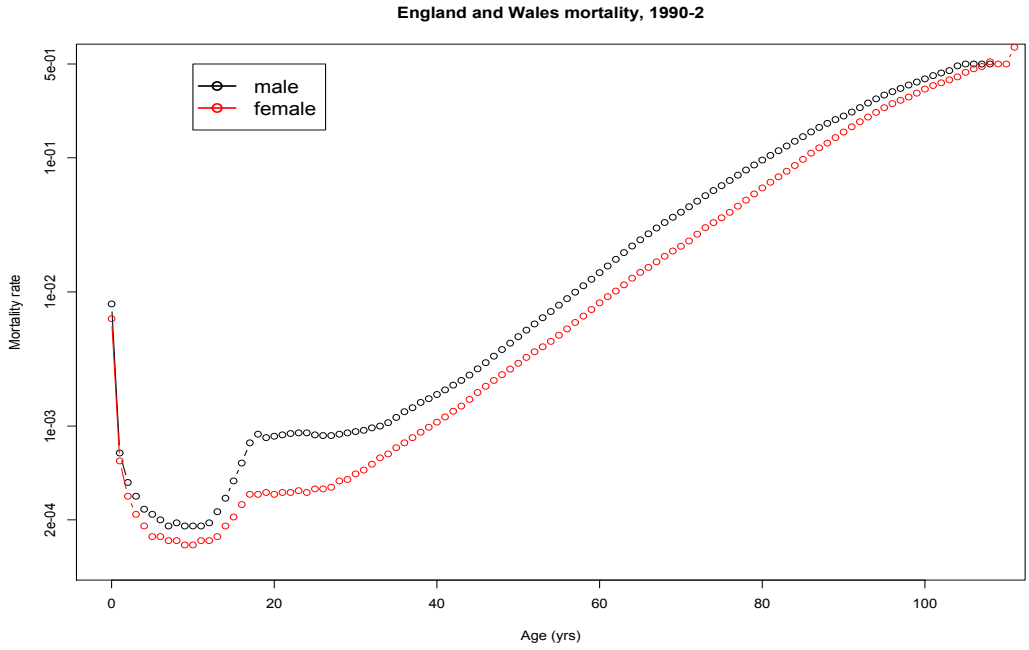


Figure 5.1: Age-specific mortality rates in England and Wales, 1990–2.

■

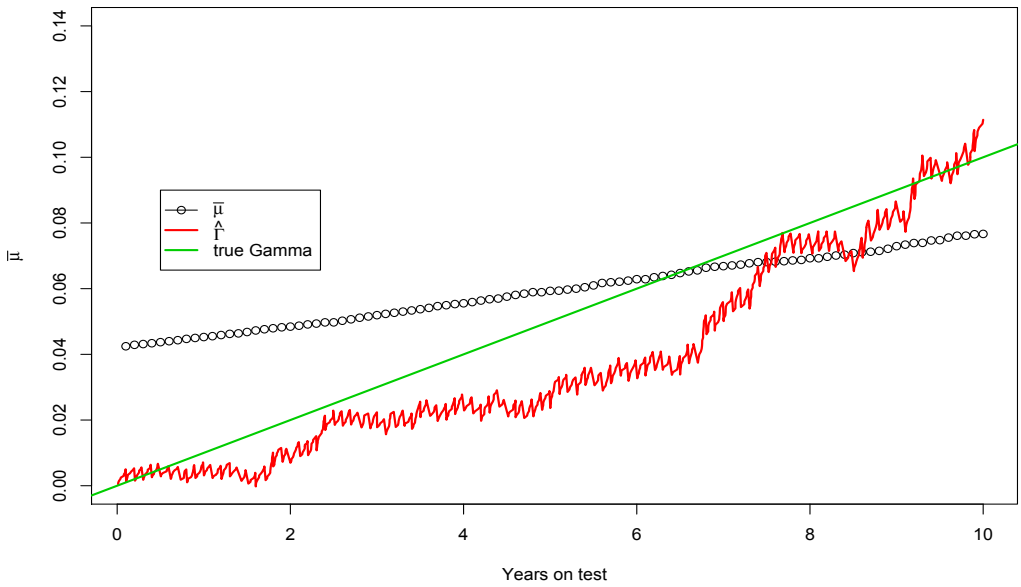


Figure 5.2: Estimating excess mortality. This combines simulated survival data with the mortality rates presented in Figure 5.1.

5.2 Excess mortality: Two-sample case

Example 5.2: NHANES survival data

Figure 5.3 shows male and female survival curves as a function of age, calculated from 16,995 subjects in the NHANES third wave (1988–94), after up to 12 years of follow-up. The subjects are divided into 4 ethnic categories (white, black, Mexican, and other). Suppose we wish to estimate the difference in survival between males and females in this study population, correcting for differences in proportion of the different ethnic groups.

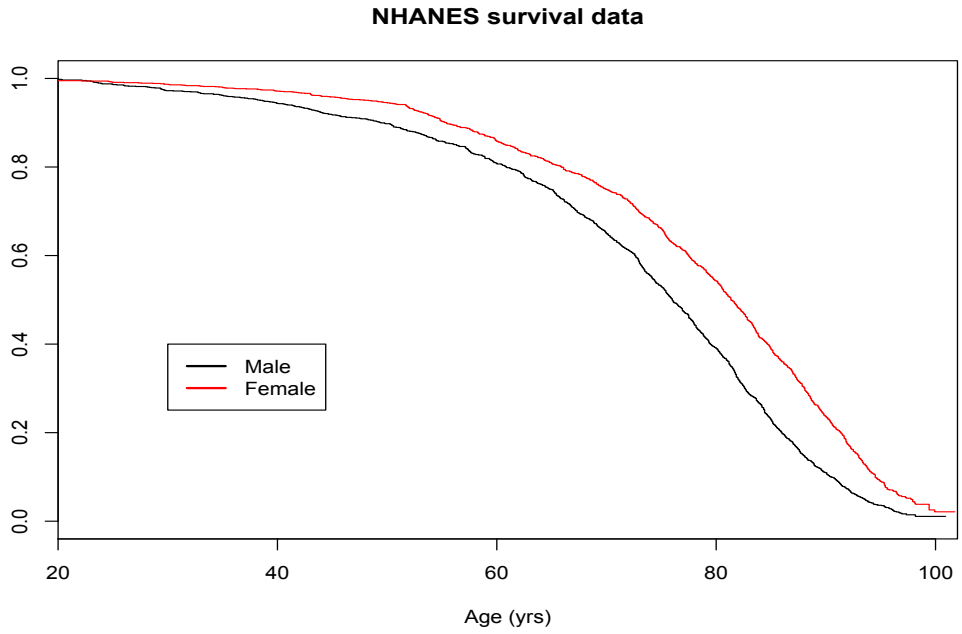


Figure 5.3: Estimated survival function from 16,995 subjects in the NHANES 3rd wave

■

We have survival data on individuals i, categorised in two different ways: There is a nuisance categorisation c_i (which might have several classes)

and a binary categorisation of interest G_i which is 0 or 1. We assume the multiplicative-intensity model, with hazard rates $\alpha(c_i;t) + G_i\gamma(t)$. In the NHANES example, G_i is sex and c_i is ethnicity.

We have the intensity for individual i of

$$\lambda_i(t) = \big(\gamma(t) + \alpha(c_i;t)\big)Y_i(t) \text{ if } G_i = 1,$$
$$\lambda_i(t) = \alpha(c_i;t)Y_i(t) \text{ if } G_i = 0.$$

For a possible category c and G =0 or 1, define

$$Y(c,G;t) := \sum_{\substack{i\,:\,c_i=c,\\ G_i=G}} Y_i(t), \qquad N(c,G;t) := \sum_{\substack{i\,:\,c_i=c,\\ G_i=G}} N_i(t).$$

Define $Y(c,-;t) := \min\{Y(c,0;t), Y(c,1;t)\}$, and

$$k_c(t) := \frac{Y(c,-;t)}{\sum_{c'} Y(c',-;t)}. \tag{5.4}$$

Then

$$\begin{aligned}\hat{\Gamma}(t) &= \sum_c \int_0^t k_c(s)\left(\frac{dN(c,1;s)}{Y(c,1;s)} - \frac{dN(c,0;s)}{Y(c,0;s)}\right)\\ &= \sum_{t_i\le t} k_c(t_i)\left(\frac{G_i}{Y(c_i,1;t_i)} - \frac{1-G_i}{Y(c_i,0;t_i)}\right)\end{aligned} \tag{5.5}$$

is an unbiased estimator for $\Gamma^*(t)$. Assuming no ties, we have

$$\mathrm{Var}\big(\hat{\Gamma}(t)\big) \le \mathbb{E}\Big[\sum_{t_i\le t}\Big(\sum_c Y(c,-;t_i)\Big)^{-2}\Big], \tag{5.6}$$

which we approximate, as usual, by the realised value

$$\sum_{t_i\le t}\Big(\sum_c Y(c,-;t_i)\Big)^{-2}. \tag{5.7}$$

The derivation of these facts is left as an exercise. As we point out

5.3 Nonparametric tests for equality: One-sample setting

We have now described methods for estimating the difference between hazard rates in different settings. Once we have an estimator with variance, it is

straightforward to turn it into a significance test. In the remainder of this chapter we describe the standard nonparametric tests for equality of survival distributions.

The simplest situation is when we are comparing observations to a given (either theoretically or by a large quantity of prior data) survival distribution. Following the approach in section 5.1 we suppose we have a null hypothesis

$$H_0: \text{ the intensity for individual } i \text{ is } \lambda_i(t) = \mu_i(t)Y_i(t).$$

We wish to test this null hypothesis against "non-crossing" alternatives, which may be understood as $\lambda_i(t) = (\mu_i(t) + \gamma_i(t))Y_i(t)$ where $\gamma_i(t)$ are all of the same sign for all times t. As before, we define

$$\bar{\mu}(t) = \sum_{i=1}^{n} \mu_i(t) \frac{Y_i(t)}{Y(t)}.$$

5.3.1 No ties

If we choose any predictable weight function $W(t)$ such that $W(t) = 0$ whenever $Y(t) = 0$ (and adopt the convention $0/0 = 0$), then under the null hypothesis

$$M(t) := \int_0^t W(s) \left(\frac{dN(s)}{Y(s)} - \bar{\mu}(s) \right) ds = \sum_{t_i \le t} \frac{W(t_i)}{Y(t_i)} - \int_0^t W(s)\bar{\mu}(s)ds$$

is a martingale, with mean zero and asymptotically normal.

The change from the estimation setting is that we estimate the variance not from the sample, but from the null hypothesis. Under the null hypothesis the intensity of the counting process is $\bar{\mu}(s)Y(s)$. By Fact 3.5 we can compute the predictable variation as

$$\langle M \rangle(t) = \int_0^t \frac{W(s)^2}{Y(s)^2} \bar{\mu}(s)Y(s)ds = \int_0^t \frac{W(s)^2 \bar{\mu}(s)}{Y(s)} ds. \tag{5.8}$$

The variance is the expectation of this, which will be difficult or impossible to compute in general, since it depends, among other things, on the probability of an individual being censored or truncated, something that we generally avoid modelling. Instead, we may approximate the expected value by the observed value of this integral.

Then we have the test statistic

$$Z(t) := \left(\int_0^t \frac{W(s)^2 \bar{\mu}(s)}{Y(s)} ds \right)^{-1/2} \left(\sum_{t_i \le t} \frac{W(t_i)}{Y(t_i)} - \int_0^t W(s)\bar{\mu}(s)ds \right), \tag{5.9}$$

which under the null hypothesis should have a standard normal distribution for any fixed t. This may be used for one-sided alternatives (hazard $> \mu_i$ or hazard $< \mu_i$) or two-sided alternatives (hazard $\neq \mu_i$).

5.3.2 Weight functions and particular tests

A popular choice of weight is simply $W(t) = Y(t)$. The resulting test is called the *log-rank test*. In this special case,

$$Z(t) = \Big(\int_0^t Y(s)\bar{\mu}(s)ds\Big)^{-1/2}\Big(N(t) - \int_0^t Y(s)\bar{\mu}(s)ds\Big).$$

Note that the variance (which is also identical with the expectation term in parentheses) is equal to the sum of cumulative hazards of all the individual null-hypothesis hazards, over the time that the individual was at risk.

Another popular class of weight functions is the Harrington-Fleming family

$$W_{HF}(t) := Y(t)S(t)^p(1 - S(t))^q, \tag{5.10}$$

for nonnegative p, q, where $S(t)$ is the survival probability under the null hypothesis. When $p = q = 0$ this is just the log-rank test. Larger values of p and q reduce the effect of deviations early and/or late in the process.

The case $p = 1, q = 0$ is essentially the popular *Peto test* (also known as the *Peto–Peto test* or *Petos' test*, as two different Petos were involved), except for slight modifications that are intended to improve the small-sample behaviour. (They take the version $W(t) := \hat{S}(t)Y(t)/(Y(t) + 1)$.)

Note that different sources disagree on which part of the test statistic is the "weight"; we are following the convention of [ABG08]. If you look in the book of [KM03] you will find the weights for the log-rank test given as $W(t) = 1$, as the factor $Y(t)$ is simply absorbed into weight. We adopt the convention here in order to emphasise the connection between the "base case" and the nonparametric estimators.

5.3.3 With ties

When tied observations are allowed, if we assume that these are merely "rounding ties" then the exact form of the test statistic will depend on how the weight function changes as the population is decremented, and it could actually depend on the order in which the events occur. Partly for this reason, it is conventional to perform significance tests as though the times were genuinely discrete, and the ties are real ties. In the one-sample case, where we have d_i events occurring at ordered times t_i, this is still quite

straightforward. We also need the discrete hazard $h_i := 1 - S(t_i)/S(t_{i-1})$, which is the probability of an event at time t_i, given survival up to time t_i.

Then

$$M(t) := \sum_{t_i \le t} \big(d_i - h_i Y(t_i)\big)$$

is a martingale that is constant except at the times t_i. As usual, we look at the weighted version

$$M^*(t) := \int_0^t \frac{W(t)}{Y(t)} dM(t) = \sum_{t_i \le t} \Big(\frac{W(t_i) d_i}{Y(t_i)} - W(t_i) h_i \Big)$$

The predictable variation is the sum of the conditional variances of the jumps. By assumption, $W(t_i), Y(t_i) \in \mathcal{F}_{t_i-}$, so

$$\langle M^* \rangle(t) = \sum_{t_i \le t} \left(\frac{W(t_i)}{Y(t_i)} \right)^2 \mathrm{Var}\big(d_i \,\big|\, Y(t_i)\big).$$

In this case, the model is that conditioned on events up to time t_i, the number of events d_i has binomial distribution with parameters $(Y(t_i), h_i)$, with variance $Y(t_i)h_i(1-h_i)$. The predictable variation is thus

$$\langle M^* \rangle(t) = \sum_{t_i \le t} \frac{W(t_i)^2 h_i (1 - h_i)}{Y(t_i)},$$

and, as usual, we may take this as an unbiased estimator for the variance of $M^*(t)$.

5.3.4 An example

Table 5.1 presents imaginary data for men aged 90 to 95. The number at risk increases and decreases, which may reflect either left truncation or that this is a "period table", reflecting different groups of individuals at risk at different ages. Here we use the weights $W \equiv 1$.

Thus, our test statistic is

$$Z := \frac{M^*(95)}{\sqrt{\langle M^* \rangle(95)}} = 2.12$$

If we are performing a two-tailed hypothesis test at the 0.95 level, we reject values of Z with modulus > 1.96. Thus, we conclude that the mortality rate in this population is significantly different from the rate in the general population.

t_i	$Y(t_i)$	d_i	$\mu(t_i)$	h_i	excess	$M^*(t_i)$	$\langle M^* \rangle(t_i)$
90	40	10	0.202	0.183	2.684	0.067	0.004
91	35	8	0.215	0.193	1.229	0.102	0.008
92	22	4	0.236	0.210	-0.625	0.074	0.016
93	14	6	0.261	0.230	2.784	0.273	0.028
94	11	4	0.279	0.243	1.322	0.393	0.045
95	7	3	0.291	0.252	1.233	0.569	0.072

Table 5.1: Table of mortality rates for an imaginary old-people's home, with standard British male mortality given as $\mu(x)$.

5.4 Non-parametric tests for equality: Two-sample setting

5.4.1 No ties

Assume we have two different groups, where the individuals have hazard rate $\alpha_0(t)$ and $\alpha_1(t)$ when at risk. Let $w(t)$ be any predictable weights, assumed to be such that $w(t) = 0$ whenever $Y_0(t)Y_1(t) = 0$ (that is, when either of the groups has no one left at risk); we adopt the convention $0/0 = 0$. We have for each group $g = 0, 1$ the Nelson–Aalen estimator

$$\hat{A}_g(t) = \int_0^t \mathbf{1}_{\{Y_g>0\}} \frac{dN_g(s)}{Y_g(s)} = \sum_{t_i^{(g)} \le t} \frac{\mathbf{1}_{\{Y_g>0\}}}{Y_g(t_i)},$$

where $t_i^{(g)}$ are the times of events belonging to individuals of group g. As before, we write

$$A_g^*(t) := \int_0^t \mathbf{1}_{\{Y_g>0\}} \alpha_g(s) ds.$$

We know that for $g = 0, 1$,

$$M_g(t) = \hat{A}_g(t) - A_g^*(t)$$

is a mean-zero martingale for each g. Thus,

$$M(t) := \int_0^t w(s)\big(dM_1(s) - dM_0(s)\big)$$

is also a mean-zero martingale. We have

$$\begin{aligned} M(t) &= \int_0^t w(s)\big(d\hat{A}_1(s) - d\hat{A}_0(s)\big) + \int_0^t w(s)\big(\mathbf{1}_{\{Y_1>0\}}\alpha_1(s) - \mathbf{1}_{\{Y_0>0\}}\alpha_0(s)\big)ds \\ &= \int_0^t w(s)\big(d\hat{A}_1(s) - d\hat{A}_0(s)\big) + \int_0^t w(s)\big(\alpha_1(s) - \alpha_0(s)\big)ds. \end{aligned}$$

The last step follows from the assumption that $w(s) = 0$ whenever either of the indicators is 0. Under the null hypothesis, then,

$$M(t) = \int_0^t w(s)\big(d\hat{A}_1(s) - d\hat{A}_0(s)\big) = \sum_{t_j \le t} \frac{(-1)^{G_j+1} w(t_j)}{Y_{G_j}(t_j)},$$

where G_i is the index of the group whose event occurs at time t_i.

The predictable variation of M is the sum of the predictable variations of M_1 and M_0 (since they are independent). Since we have the representation

$$M_g(t) = \int_0^t \frac{w(s)}{Y_g(s)}\big(dN_g(s) - Y_g(s)\alpha(s)\big),$$

for the counting process N_g with intensity $Y_g\alpha$, we have by Fact 3.5

$$\langle M_g \rangle(t) = \int_0^t \left(\frac{w(s)}{Y_g(s)}\right)^2 Y_g(s)\alpha(s)ds.$$

Thus, under the null hypothesis

$$\begin{aligned} \langle M \rangle(t) &= \int_0^t \left(\frac{w(s)^2}{Y_0(s)} + \frac{w(s)^2}{Y_1(s)}\right)\alpha(s)ds \\ &= \int_0^t \frac{w(s)^2 Y_\cdot(s)}{Y_0(s)Y_1(s)}\alpha(s)ds \end{aligned} \tag{5.11}$$

Of course, we generally don't know α; so we replace $\alpha(s)ds$ by the estimator $d\hat{A}(s) = dN_\cdot(s)/Y_\cdot(s)$ (where the dot represents summation over the index). Thus we obtain

$$\begin{aligned} \langle M \rangle(t) &\approx \int_0^t \frac{w(s)^2 Y_\cdot(s)}{Y_0(s)Y_1(s)} d\hat{A}(s)ds \\ &= \sum_{t_i \le t} \frac{w(t_i)^2}{Y_0(t_i)Y_1(t_i)}. \end{aligned} \tag{5.12}$$

We take this as an estimate for the variance

$$\mathrm{Var}(M(t)) \approx \sum_{t_i \le t} \frac{w(t_i)^2}{Y_0(t_i)Y_1(t_i)}. \tag{5.13}$$

5.4.2 Weight functions and particular tests

The weight functions look slightly different in the 2-sample case. The 2-sample log rank test has weights

$$w_{LR}(t) = \frac{Y_0(t)Y_1(t)}{Y_\cdot(t)}. \tag{5.14}$$

The log rank test statistic is then

$$Z(t) = \Big(\sum_{t_j \le t} \frac{Y_0(t_j)Y_1(t_j)}{Y_\cdot(t_j)^2}\Big)^{-1/2}\Big(N_1(t) - \sum_{t_j \le t} \frac{Y_1(t_j)}{Y_\cdot(t_j)}\Big). \tag{5.15}$$

Note that the variance (which is also identical with the expectation term in parentheses) is equal to the sum of cumulative hazards of all the individual null-hypothesis hazards, over the time that the individual was at risk.

Other weight functions are defined similarly. The Harrington-Fleming family is

$$w_{HF}(t) := \hat{S}(t-)^p(1 - \hat{S}(t-))^q \frac{Y_0(t)Y_1(t)}{Y_\cdot(t)}, \tag{5.16}$$

for nonnegative p, q, where $\hat{S}(t)$ is an estimator for the survival probability under the null hypothesis. Since this states that the two populations have the same survival distribution, $\hat{S}$ is just an estimator — for example, the Kaplan–Meier estimator — for survival, where we treat all the individuals as coming from a single population.

When $p = q = 0$ this is just the log-rank test. Larger values of p and q reduce the effect of deviations early and/or late in the process. The Peto test is essentially the Harrington-Fleming test with parameters $(1, 0)$, except for the small modification to

$$w_{HF}(t) := \hat{S}(t-)\frac{Y_0(t)Y_1(t)}{Y_\cdot(t) + 1}, \tag{5.17}$$

Since the log-rank weights are already fairly complicated, [ABG08] simplifies matters by defining the weight to be

$$w(t) = K(t)\frac{Y_0(t)Y_1(t)}{Y_\cdot(t)}.$$

This way me may think of just the weights K, which are now constant for the log-rank test, and much simpler for many other standard tests as well.

The test statistic is

$$Z(t) := \Big(\sum_{t_i \le t} \frac{K(t_i)^2 Y_0(t_i)Y_1(t_i)}{Y_\cdot(t_i)^2}\Big)^{-1/2} \sum_{t_i \le t} \frac{(-1)^{G_i+1}K(t_i)Y_{1-G_i}(t_i)}{Y_\cdot(t_i)}.$$

5.4.3 With ties

As in section 5.3.3, we assume that the ties are genuine, so that we must treat the survival curve as having discrete hazards $h_j := 1 - S(t_j)/S(t_j-)$, which is the probability of an event at time t_j, given survival up to time t_j. We write $d_j^{(g)}$ $(g = 0, 1)$ for the number of events of type g at time t_j and $d_j = d_j^{(1)} + d_j^{(0)}$. (Note: We will only be needing to refer to the hazards h_j under the null hypothesis, hence no need to define separate hazards for the two groups.)

Assuming the null hypothesis, the d_j individuals having events at time t_j are uniformly chosen from the $Y_\cdot(t_j)$ individuals at risk at time t_j. This is like the variance of the number of red balls obtained when drawing d_j at random from an urn containing $Y_0(t_j)$ red and $Y_1(t_j)$ blue balls. Thus

$$\mathbb{E}\left[d_j^{(1)} - d_j \frac{Y_1(t_j)}{Y_\cdot(t_j)} \,\middle|\, d_j\right] = 0,$$

so that

$$\mathbb{E}\left[d_j^{(1)} - d_j \frac{Y_1(t_j)}{Y_\cdot(t_j)} \middle| \mathcal{F}_{t_j-}\right] = \mathbb{E}\left[\mathbb{E}\left[d_j^{(1)} - d_j \frac{Y_1(t_j)}{Y_\cdot(t_j)} \,\middle|\, d_j\right] \middle| \mathcal{F}_{t_j-}\right] = 0.$$

Thus

$$M(t) := \sum_{t_j \le t} K(t_j)\left(d_j^{(1)} - d_j \frac{Y_1(t_j)}{Y_\cdot(t_j)}\right) = \sum_{t_j \le t} w(t_j) \frac{Y_\cdot(t_j)}{Y_0(t_j)Y_1(t_j)}\left(d_j^{(1)} - d_j \frac{Y_1(t_j)}{Y_\cdot(t_j)}\right)$$
$$= \sum_{t_j \le t} w(t_j)\left(\frac{d_j^{(1)}}{Y_1(t_j)} - \frac{d_j^{(0)}}{Y_0(t_j)}\right).$$

is a martingale that is constant except at the times t_j.[1]

The predictable variation is the sum of the conditional variances of the jumps. By assumption, $w(t_j), Y(t_j) \in \mathcal{F}_{t_j-}$, so

$$\langle M \rangle(t) = \sum_{t_j \le t} K(t_j)^2 \operatorname{Var}\left(d_j^{(1)} \,\middle|\, \mathcal{F}_{t_j-}\right).$$

Unlike the corresponding calculation in section 5.4.3, in this one we have no specific hazard rate given by the null hypothesis.

[1] As discussed in section 5.3.3, the weight functions w and K are completely equivalent, and the choice of one or the other is purely a matter of convenience. Depending on context, one or the other may appear more natural.

Let $\mathcal{G}_j = \mathcal{F}_{t_j-} \vee \langle d_j \rangle$, where $d_j := d_j^{(1)} + d_j^{(0)}$. (So it represents the information available before time t_j, together with d_j, and so is intermediate between $\mathcal{F}_{t_j}$ and $\mathcal{F}_{t_j-}$.) We may do the conditioning in two steps, first conditioning on $\mathcal{G}_j$ — effectively, conditioning on $Y_0(t_j)$, $Y_1(t_j)$, and d_j, and then on the smaller σ-algebra $\mathcal{F}_{t_j-}$.

$$\operatorname{Var}(d_j^{(1)} \mid \mathcal{F}_{t_j-}) = \mathbb{E}\left[\operatorname{Var}\big(d_j^{(1)} \mid \mathcal{G}_j\big) \mid \mathcal{F}_{t_j-}\right]. \tag{5.18}$$

Because the sampling is without replacement, the distribution is not binomial (though obviously close when Y is large), but *hypergeometric*, so

$$\operatorname{Var}\big(d_j^{(1)} \mid \mathcal{G}_j\big) = \frac{d_j(Y_\cdot(t_j) - d_j)Y_0(t_j)Y_1(t_j)}{Y_\cdot(t_j)^2(Y_\cdot(t_j) - 1)}.$$

This gives us the increment to the variation

$$K(t_j)^2 \frac{\mathbb{E}[d_j(Y_\cdot(t_j) - d_j)|\mathcal{F}_{t_j-}]Y_0(t_j)Y_1(t_j)}{Y_\cdot(t_j)^2(Y_\cdot(t_j) - 1)} = w(t_j)^2 \frac{\mathbb{E}[d_j(Y_\cdot(t_j) - d_j)|\mathcal{F}_{t_j-}]}{Y_0(t_j)Y_1(t_j)(Y_\cdot(t_j) - 1)}.$$

The true variance is given by the expected value of the predictable variation, but as usual, we take the realised value

$$\hat{\sigma}^2 = \sum_{t_j \le t} w(t_j)^2 \frac{d_j\left(Y_\cdot(t_j) - d_j\right)}{Y_0(t_j)Y_1(t_j)(Y_\cdot(t_j) - 1)} \tag{5.19}$$

as an unbiased estimator for the variance.

We have, for sufficiently large sample sizes, that the test statistic

$$\left(\sum_{t_j \le t} w(t_j)^2 \frac{d_j(Y_\cdot(t_j) - d_j)}{Y_0(t_j)Y_1(t_j)(Y_\cdot(t_j) - 1)}\right)^{-1/2} \left(\sum_{t_j \le t} w(t_j)\left(d_j^{(1)} - d_j \frac{Y_1(t_j)}{Y_\cdot(t_j)}\right)\right) \tag{5.20}$$

has approximately standard normal distribution.

5.4.4 The AML example

In the 1970s it was known that individuals who had gone into remission after chemotherapy for acute lymphatic leukemia would benefit — by longer remission times — from a course of continuing "maintenance" chemotherapy. A study [EEH+77] pointed out that "Despite a lack of conclusive evidence, it has been assumed that maintenance chemotherapy is useful in the management of acute myelogenous leukemia (AML)." The study set out to test this

assumption, comparing the duration of remission between an experimental group that received the additional chemotherapy, and a control group that did not. (This analysis is based on the discussion in [MGM01].) We will analyse these data in various ways in this lecture.

The data are from a preliminary analysis of the data, before completion of the study. The duration of complete remission in weeks was given for each patient (11 maintained, 12 non-maintained controls); those who were still in remission at the time of the analysis are censored observations. The data are given in Table 5.2. They are included in the `survival` package of `R`, under the name `aml`.

Table 5.2: Times of complete remission for preliminary analysis of AML data, in weeks. Censored observations denoted by +.

maintained	9 13 13+ 18 23 28+ 31 34 45+ 48 161+
non-maintained	5 5 8 8 12 16+ 23 27 30 33 43 45

The first thing we do is to estimate the survival curves. The summary data and computations are given in Table 5.3. The Kaplan–Meier survival curves are shown in Figure 5.4. In Table 5.4 we show the computations for confidence intervals just for the Kaplan–Meier curve of the maintenance group. The confidence intervals are based on the logarithm of survival. That is, the bounds on the confidence interval are

$$\exp\left\{\log \hat{S}(t) \pm z\sqrt{\sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}}\right\},$$

where z is the appropriate quantile of the normal distribution.

We could also use

$$\hat{S}(t) \pm z\hat{S}(t)\sqrt{\sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}}.$$

Note that the approximation cannot be assumed to be very good in this case, since the number of individuals at risk is too small for the asymptotics to be reliable. We show the confidence intervals in Figure 5.5.

Important: The estimate of the variance is more generally reliable than the assumption of normality, particularly for small numbers of events. Thus,

Table 5.3: Computations for the Kaplan–Meier and Nelson–Aalen survival curve estimates of the AML data.

	Maintenance						Non-Maintenance (control)					
t_i	$Y(t_i)$	d_i	$\hat{h}_i$	$\hat{S}(t_i)$	$\hat{A}(t_i)$	$\widetilde{S}(t_i)$	$Y(t_i)$	d_i	$\hat{h}_i$	$\hat{S}(t_i)$	$\hat{A}(t_i)$	$\widetilde{S}(t_i)$
5	11	0	0.00	1.00	0.00	1.00	12	2	0.17	0.83	0.17	0.85
8	11	0	0.00	1.00	0.00	1.00	10	2	0.20	0.67	0.37	0.69
9	11	1	0.09	0.91	0.09	0.91	8	0	0.00	0.67	0.37	0.69
12	10	0	0.00	0.91	0.09	0.91	8	1	0.12	0.58	0.49	0.61
13	10	1	0.10	0.82	0.19	0.83	7	0	0.00	0.58	0.49	0.61
18	8	1	0.12	0.72	0.32	0.73	6	0	0.00	0.58	0.49	0.61
23	7	1	0.14	0.61	0.46	0.63	6	1	0.17	0.49	0.66	0.52
27	6	0	0.00	0.61	0.46	0.63	5	1	0.20	0.39	0.86	0.42
30	5	0	0.00	0.61	0.46	0.63	4	1	0.25	0.29	1.11	0.33
31	5	1	0.20	0.49	0.66	0.52	3	0	0.00	0.29	1.11	0.33
33	4	0	0.00	0.49	0.66	0.52	3	1	0.33	0.19	1.44	0.24
34	4	1	0.25	0.37	0.91	0.40	2	0	0.00	0.19	1.44	0.24
43	3	0	0.00	0.37	0.91	0.40	2	1	0.50	0.10	1.94	0.14
45	3	0	0.00	0.37	0.91	0.40	1	1	1.00	0.00	2.94	0.05
48	2	1	0.50	0.18	1.41	0.24	0	0				

Table 5.4: Variance estimates for cumulative hazard of the maintenance population in the AML data. "Lower" and "upper" are bounds for 95% confidence intervals.

t_i	$Y(t_i)$	d_i	$\frac{1}{Y(t_i)^2}$	$\sigma^2(t_i)$	lower	upper
9	11	1	0.008	0.008	0.000	0.269
13	10	1	0.010	0.018	0.000	0.456
18	8	1	0.016	0.034	0.000	0.677
23	7	1	0.020	0.054	0.002	0.915
31	5	1	0.040	0.094	0.057	1.26
34	4	1	0.062	0.157	0.133	1.68
48	2	1	0.25	0.407	0.159	2.66

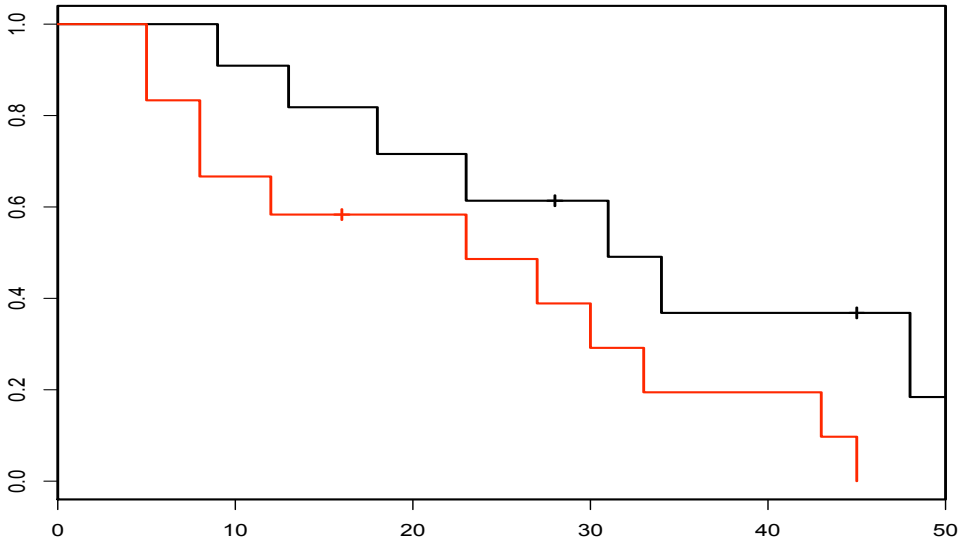


Figure 5.4: Kaplan–Meier estimates of survival in maintenance (black) and non-maintenance groups in the AML study.

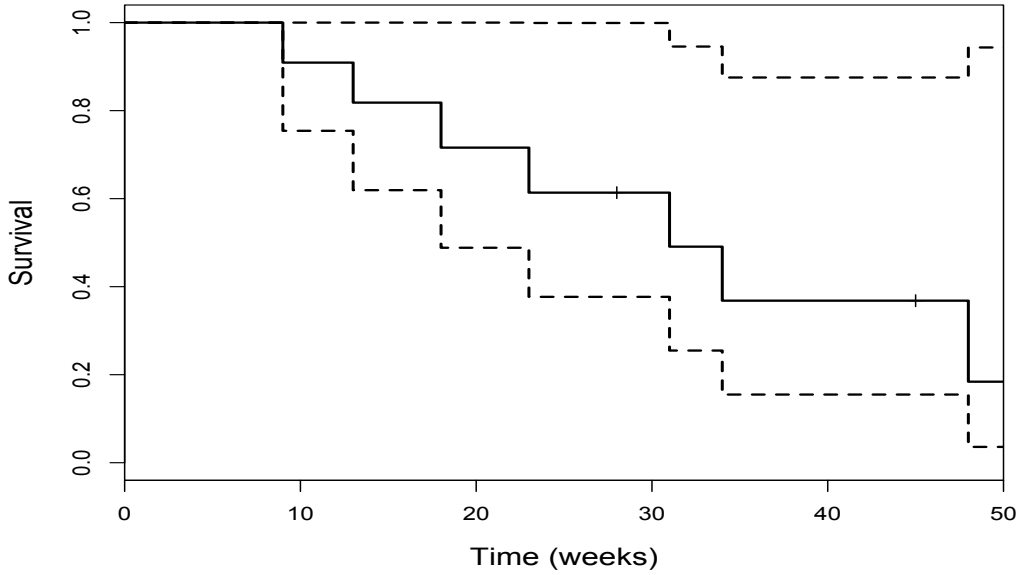


Figure 5.5: Estimated of 95% confidence intervals for survival in maintenance group of the AML study.

the first line in Table 5.4 indicates that the estimate of $\hat{A}(9)$ is associated with a variance of 0.008. The error in this estimate is on the order of $Y(t_i)^{-3}$, so it's potentially about 10% of the indicated value. On the other hand, the number of events observed has binomial distribution, with parameters around $(11, 0.9)$, so it's very far from a normal distribution.

We can use these tests to compare the survival of the two groups. The relevant quantities are tabulated in Table 5.5. The column σ_j^2 gives the increments to the approximate variance

$$\sigma_j^2 = \frac{d_j Y_0(t_j) Y_1(t_j)(Y.(t_j) - d_j)}{Y.(t_j)^2 (Y.(t_j) - 1)}$$

t_i	$Y_0(t_i)$	$Y_1(t_i)$	$d_0(t_i)$	$d_1(t_i)$	σ_i^2	Peto wt.	H–F $(0,1)$ wt.
5	11	12	0	2	0.476	0.958	0.000
8	11	10	0	2	0.474	0.875	0.083
9	11	8	1	0	0.244	0.792	0.167
12	10	8	0	1	0.247	0.750	0.208
13	10	7	1	0	0.242	0.708	0.250
18	8	6	1	0	0.245	0.661	0.292
23	7	6	1	1	0.456	0.614	0.339
27	6	5	0	1	0.248	0.519	0.433
30	5	4	0	1	0.247	0.467	0.481
31	5	3	1	0	0.234	0.416	0.533
33	4	3	0	1	0.245	0.364	0.584
34	4	2	1	0	0.222	0.312	0.636
43	3	2	0	1	0.240	0.260	0.688
45	3	1	0	1	0.188	0.208	0.740
48	2	0	1	0	0.000	0.139	0.792

Table 5.5: Data for testing equality of survival in AML experiment.

When the weights are all taken equal, we compute $Z = -1.84$, whereas the Peto weights — which reduce the influence of later observations — give us $Z = -1.67$. This yields one-sided p-values of 0.033 and 0.048 respectively — a marginally significant difference — or two-sided p-values of 0.065 and 0.096.

5.4.5 Kidney dialysis example

This is example 7.2 from [KM03]. The data are from a clinical trial of alternative methods of placing catheters in kidney dialysis patients. The event time is the first occurrence of an exit-site infection. The data are in the `KMsurv` package, in the object `kidney`. (Note: There is a different data set with the same name in the `survival` package. To make sure you get the right one, enter `data(kidney,package='KMsurv')`.) The Kaplan–Meier estimatoris shown in Figure 5.6. There are two survival curves, corresponding to the two different methods.

We show the calculations for the nonparametric test of equality of distributions in Table 5.6. The log-rank test — obtained by simply dividing the sum of all the deviations by the square root of the sum of terms in the σ_i^2 column — is only 1.59, so not significant. With the Peto weights the statistic is only 1.12. This is not surprising, because the survival curves are close together (and actually cross) early on. On the other hand, they diverge later, suggesting that weighting the later times more heavily would yield a significant result. It would not be responsible statistical practice to choose a different test after seeing the data. On the other hand, if we had started with the belief that the benefits of the percutaneous method are cumulative, so that it would make sense to expect the improved survival to appear later on, we might have planned from the beginning to use the Harrington–Fleming weights with, for example, $p = 0, q = 1$, tabulated in the last column. Applying these weights gives us a test statistic $Z_{FH} = 3.11$, implying a highly significant difference.

We conclude with the `R` code used to generate Figure 5.6 and Table 5.6

```
require('KMsurv')
data(kidney,package='KMsurv')
attach(kidney)
kid.surv=Surv(time,delta)
kid.fit=survfit(kid.surv~type)
plot(kid.fit,col=1:2,xlab='Time to infection (months)',
   main='Kaplan-Meier plot for kidney dialysis data')

t=sort(unique(time))
d=lapply(1:2,function(i) sapply(t,function(T) sum((time[type==i]==T)&(delta[type==i]==1))))
n=lapply(1:2,function(i) sapply(t,function(T) sum(time[type==i]>=T)))

keep=(n[[1]]*n[[2]]>0)
t=t[keep]
d=lapply(d,function(D) D[keep])
n=lapply(n,function(D) D[keep])
```

t_i	$Y_0(t_i)$	$Y_1(t_i)$	$d_0(t_i)$	$d_1(t_i)$	σ_i^2	Peto wt.	H–F $(0,1)$ wt.
0.5	43	76	0	6	1.326	0.992	0.000
1.5	43	60	1	0	0.243	0.941	0.050
2.5	42	56	0	2	0.485	0.931	0.059
3.5	40	49	1	1	0.489	0.912	0.078
4.5	36	43	2	0	0.490	0.890	0.099
5.5	33	40	1	0	0.248	0.867	0.121
6.5	31	35	0	1	0.249	0.854	0.133
8.5	25	30	2	0	0.487	0.839	0.146
9.5	22	27	1	0	0.247	0.807	0.176
10.5	20	25	1	0	0.247	0.790	0.193
11.5	18	22	1	0	0.247	0.770	0.210
15.5	11	14	1	1	0.472	0.741	0.230
16.5	10	13	1	0	0.246	0.681	0.289
18.5	9	11	1	0	0.247	0.649	0.319
23.5	4	3	1	0	0.245	0.568	0.351
26.5	2	3	1	0	0.240	0.473	0.432

Table 5.6: Data for kidney dialysis study.

```
ddot=d[[1]]+d[[2]]
ndot=n[[1]]+n[[2]]

si=ddot*n[[1]]*n[[2]]*(ndot-ddot)/ndot/ndot/(ndot-1)
petos=cumprod((ndot-ddot)/(ndot))
petow=c(1,petos[1:(length(petos)-1)])*ndot/(ndot+1)
fhw=c(0,(1-petos[1:(length(petos)-1)]))
ei=ddot*n[[1]]/ndot
wk=rep(1,length(ei))

zLR=sum(wk*(d[[1]]-ei))/sqrt(sum(wk^2*si))
zP=sum(petow*(d[[1]]-ei))/sqrt(sum(petow^2*si))
zFH=sum(fhw*(d[[1]]-ei))/sqrt(sum(fhw^2*si))

xt=xtable(cbind(t,n[[1]],n[[2]],d[[1]],d[[2]],si,petow,fhw),
          display=c('d','f','d','d','d','d','f','f','f'),
          digits=c(0,1,0,0,0,0,3,3,3))
print(xt,include.rownames=FALSE,include.colnames=FALSE)
```

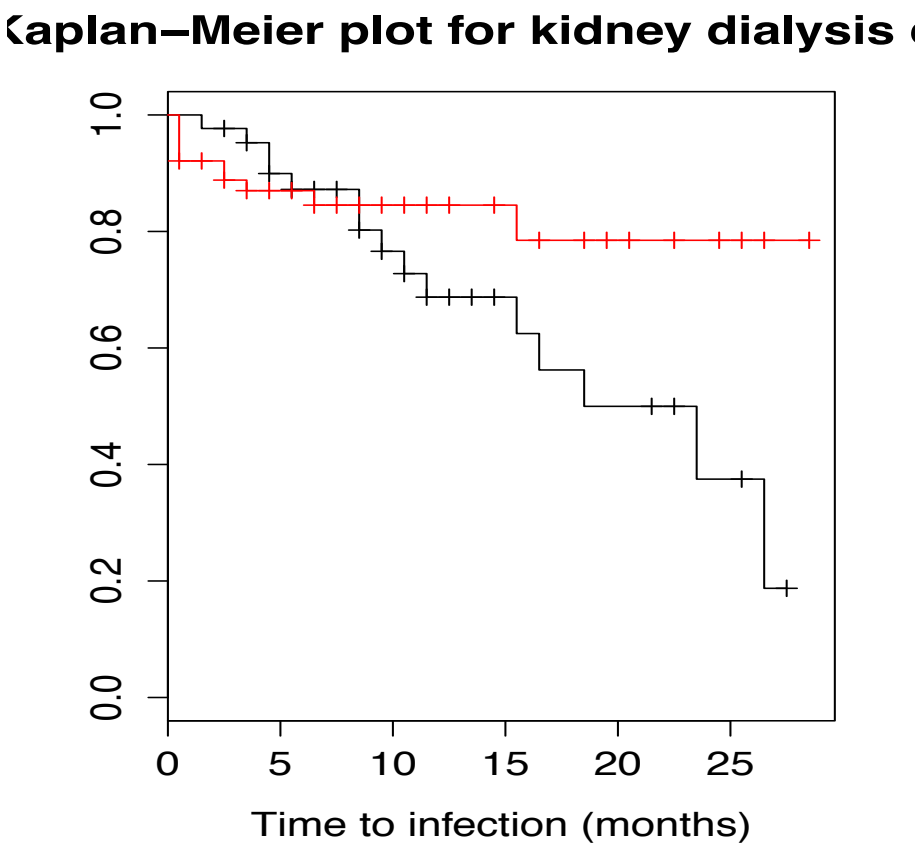


Figure 5.6: Plot of Kaplan–Meier survival curves for time to infection of dialysis patients, based on data described in section 1.4 of [KM03]. The black curve represents 43 patients with surgically-placed catheter; the red curve 76 patients with percutaneously placed catheter.