5 Rates

We have shown how, by splitting the follow-up period into small enough bands, the importance of arbitrary assumptions about when the losses occur can be minimized. We now follow this argument to its logical conclusion and divide the follow-up into infinitely small time bands.

5.1 The probability rate

As the bands get shorter, the conditional probability that a subject fails during any one band gets smaller. When a band shrinks towards a single moment of time, the conditional probability of failure during the band shrinks towards zero, but the conditional probability of failure *per unit time* converges to a quantity called the *probability rate.* This quantity is sometimes called the *instantaneous* probability rate to emphasize the fact that it refers to a moment in time. Other names are *hazard* rate and *force of mortality.*

The probability rate refers to an *individual subject.* This is counter-intuitive to many epidemiologists, who think of a rate as an empirical summary of the frequency of failures in a group observed over time. We show in the next section that such a summary is, in fact, the most likely value of the common probability rate for the subjects in the group. It is general practice in epidemiology to refer to both the probability rate and its estimated value as the rate, even though this leads to many logical absurdities. We have tried to keep as close as possible to this tradition, while avoiding the logical contradictions, by referring to the probability rate as the rate parameter and its estimated value as the observed rate.

5.2 Estimating the rate parameter

Even though the rate parameter refers to a single individual it is not possible to estimate its value from the experience of that individual. The estimate must be based on the experience of a group of subjects assumed to have the same rate. Similarly, even though the rate parameter refers to a single moment of time, its estimated value is usually based on a period of follow-up over which the rate is assumed to be constant. The estimated rate for this period then refers to the constant value which the rate parameter takes at all time points during the period.

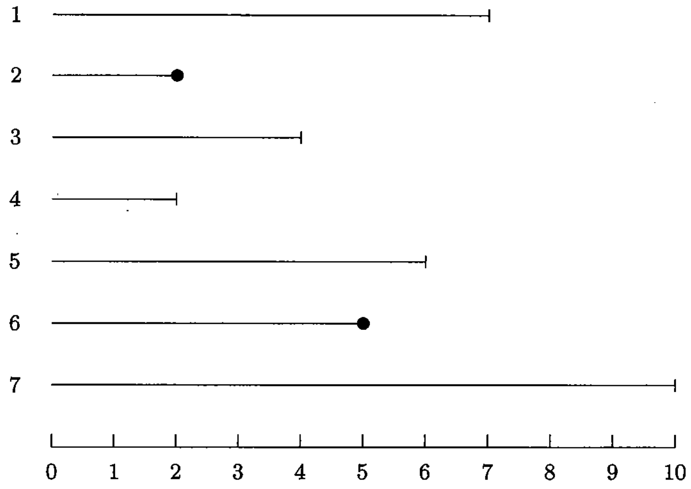

Fig. 5.1. The follow-up experience of 7 subjects.

The rate parameter over a follow-up period is estimated by dividing the period into a number of small time bands of equal length and estimating the *common* probability of failure for each of the bands. This is divided by the length of a band to get the rate per unit time. The process is illustrated using the follow-up experience of 7 subjects shown in Fig. 5.1, in which the follow-up experience of the subjects is shown as lines which end when follow-up ends. The lines for those subjects who fail end with a •, while those whose observation time is censored end with a short bar. The follow-up period has been divided into 10 short bands and for the present we shall assume that follow-up always stops at the end of a short band. From the figure we see that the follow-up of subject 1 stops after 7 bands due to censoring. For subject 6 the follow-up stops after 5 bands when the subject fails, and so on.

Exercise 5.1. How many observations of one subject through one time band are observed? How many of these ended in failure?

Assuming that the rate parameter is constant over the follow-up period, the conditional probability of failure is the same for all bands and its most likely value is 2/36. The most likely value of the corresponding rate parameter is 2/36 divided by the length of the bands. Suppose for illustration that each band has length 0.05 years. The most likely value of the rate parameter is then

$$\frac{2}{(36 \times 0.05)} = 1.11 \text{ per year.}$$

Note that 36 × 0.05, which equals 1.8 years, is the total observation time for the 7 subjects.

Now suppose that five times as many bands are used, so that each is 0.01 years in length. The most likely value of the probability of failure for these bands is 2/180, but the most likely value of the corresponding rate stays the same because there are now 180 bands of length 0.01 years and 180×0.01 is the same as 36×0.05, both being equal to the total observation time, added over subjects. In general, then, as the bands shrink to zero, the most likely value of the rate parameter is

$$\frac{\text{Total number of failures}}{\text{Total observation time}}.$$

Note that assumption that events occur at the end of bands is automatically true when the bands shrink to zero. This mathematical device of dividing the time scale into shorter and shorter bands is used frequently in this book, and we have found it useful to introduce the term *clicks* to describe these very short time bands.

Time can be measured in any convenient units, so that a rate of 1.11 per year is the same as a rate of 11.1 per 10 years, and so on. The total observation time added over subjects is known in epidemiology as the *person-time* of observation and is most commonly expressed as person-years. Because of the way they are calculated, estimates of rates are often given the units *per person-year* or *per 1000 person-years.*

The use of the general formula for the estimated value of a rate is now illustrated using data from a computer simulation of 30 subjects who are liable to only one disease (the failure) and the follow-up is indefinitely long, so that eventually all subjects develop the disease. The only variable in the outcome is how long it takes for the disease to develop, and these times are shown in Table 5.1.

Exercise 5.2. Using the time interval from the start of the study to the moment when the last subject develops the disease, find the total observation time for the 30 subjects and hence estimate the rate for this interval. Give your answer per 10^3 person-years as well.

Exercise 5.3. The previous exercise is rather unrealistic. Real follow-up studies are of limited duration and not all of the subjects will fail during the study period. Estimate the rate from a study in which the same subjects are observed only for the first five years.

Table 5.1. Time until the disease develops, for 30 subjects

Subject	Years	Subject	Years
1	19.6	16	0.6
2	10.8	17	2.1
3	14.1	18	0.8
4	3.5	19	8.9
5	4.8	20	11.6
6	4.6	21	1.3
7	12.2	22	3.4
8	14.0	23	15.3
9	3.8	24	8.5
10	12.6	25	21.5
11	12.8	26	8.3
12	12.1	27	0.4
13	4.7	28	36.5
14	3.2	29	1.1
15	7.3	30	1.5

5.3 The likelihood for a rate

The argument of the last section, although leading to the most likely value of the rate parameter, does not allow us to explore the support for other values. In this section we shall obtain a formula for the likelihood for a rate parameter.

Consider a more general example in which D failures are observed for a total of N clicks of time, each of duration h years, where h is very small and N is very large. The total observation time in years is $Y = Nh$. Let π be the conditional probability of failure during a click. Then the likelihood for π is

$$(\pi)^D(1-\pi)^{N-D}.$$

Let the corresponding rate parameter be λ, where, because h is small,

$$\lambda = \pi/h.$$

The likelihood for λ follows by replacing π by λh, and is

$$(\lambda h)^D(1-\lambda h)^{N-D}.$$

The log likelihood for λ is therefore

$$D\log(\lambda) + D\log(h) + (N-D)\log(1-\lambda h).$$

To see what happens when time is truly continuous, consider the behaviour of this expression as the click duration, h, becomes progressively shorter. Since the total observation time Y remains unchanged it follows that the number of clicks, N, must become progressively larger. As h becomes smaller and N becomes larger, eventually $N - D$ becomes nearly the same as N, and λh becomes so small that

$$\log(1 - \lambda h) \approx -\lambda h.$$

(This property of the logarithmic function is discussed in Appendix A.) Making these substitutions, the log likelihood becomes

$$D\log(\lambda) + D\log(h) - N\lambda h.$$

The term $D\log(h)$ does not depend on λ and is irrelevant since it cancels out in log likelihood ratios. Omitting this term and noting that Nh is the total observation time, Y, we obtain the following simplified expression for the log likelihood:

$$D\log(\lambda) - \lambda Y.$$

The corresponding likelihood,

$$(\lambda)^D \exp(-\lambda Y),$$

is called the *Poisson likelihood* after the French mathematician. As we would expect from the previous section it takes its maximum value when $\lambda = D/Y$.

To illustrate the use of this likelihood, suppose 7 cases are observed and the total observation time is 500 person-years. Then the log likelihood for λ is

$$7\log(\lambda) - 500\lambda.$$

A graph of the log likelihood ratio versus λ is shown in Fig. 5.2. The maximum value of the log likelihood occurs at

$$\lambda = 7/500 = 0.014 \text{ per person-year.}$$

The supported range for λ may be found from the graph by reading off the values of λ at which the log likelihood ratio has reduced to -1.353. In this case the graph shows that the supported range for λ is from 7.0×10^{-3} to 24.6×10^{-3} per person-year.

Exercise 5.4. Calculate the value of the log likelihood at $\lambda = 0.01$, $\lambda = 0.014$, and $\lambda = 0.02$. Using the fact that the log likelihood is at its maximum when $\lambda = 0.014$ calculate the log likelihood ratio for $\lambda = 0.01$ and $\lambda = 0.02$.

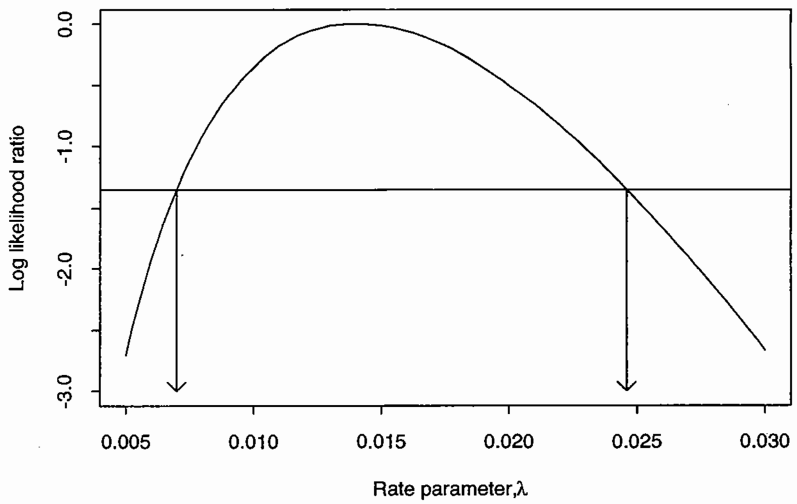

Fig. 5.2. Log likelihood ratio for λ.

If we wish to estimate the rate over a restricted period of observation the argument requires only trivial modification; only the person-clicks falling in the period of interest contribute information so that D and Y refer to the number of events and the observation time which occur within the period.

5.4 Cumulative survival probability in terms of the rate

Suppose a subject experiences a constant rate λ with no possibility of loss during the follow-up. The cumulative probability that he or she will survive a given period of time, T, may be found from λ by dividing the period into N clicks, each of length h, so that $T = Nh$. The conditional probability of failure at each click is λh, so that the probability of surviving N such clicks is

$$(1 - \lambda h)^N.$$

The log of this cumulative survival probability is

$$N\log(1 - \lambda h)$$

and since $\log(1 - \lambda h)$ may be replaced by $-\lambda h$ when h is small this becomes

$$-\lambda Nh = -\lambda T.$$

The quantity λT is called the *cumulative failure rate.* With this terminology we have the fundamental result that

$$\log(\text{Cumulative survival probability}) = -\text{Cumulative failure rate}$$

Applying the antilog function, exp(), to both sides of this relationship yields the alternative form:

$$\begin{aligned}\text{Cumulative survival probability} &= \exp(-\text{Cumulative failure rate})\\ &= \exp(-\lambda T).\end{aligned}$$

Exercise 5.5. Using your estimate of the rate for the 30 subjects shown in Table 5.1 (Exercise 5.2), calculate the probability of survival for the first 5 years, and hence the 5-year risk. Compare this with the proportion of subjects observed to fail in this period (see Exercise 5.3).

An important special case concerns *rare events*, in which the cumulative survival is large and the cumulative risk is small. Since $\log(1-x) \approx -x$ when x is small,

$$\begin{aligned}\log(\text{Cumulative survival probability}) &= \log(1-\text{Cumulative risk})\\ &\approx -\text{Cumulative risk},\end{aligned}$$

so the cumulative risk and the cumulative failure rate are approximately equal for rare events.

5.5 Rates that vary with time

We have assumed that the rate parameter is constant over the follow-up period and this may be unrealistic over an extended follow-up. However, provided the rate parameter is not changing too quickly, the follow-up period can be divided into broad bands during which the rate can be assumed to be constant. This implies abrupt changes in the rate parameter from one band to the next, but even such a crude model proves useful in practice provided the changes are not too large.

Consider the first band and let D^1 be the number of failures Y^1 the total observation time and λ^1 the rate parameter. The log likelihood for λ^1 is

$$D^1 \log(\lambda^1) - \lambda^1 Y^1$$

and similarly for further bands. Thus once failures and total observation time have been partitioned between the time bands estimation of band-specific rates proceeds as before.

Exercise 5.6. Fig. 5.3 illustrates observation of three subjects across three time bands, showing the observation time (years) for each subject in each band. What are the estimated failure rates for each of the bands?

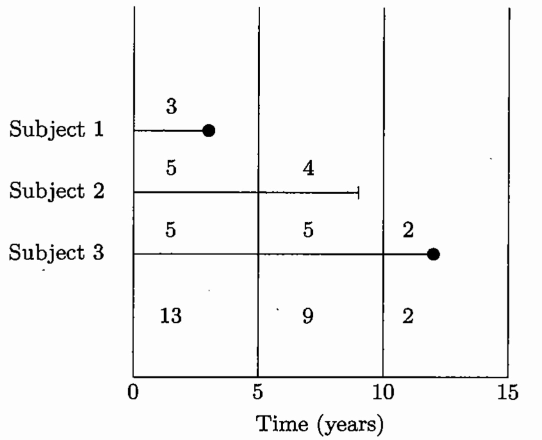

Fig. 5.3. Survival of three subjects across three time bands.

The relationship between the cumulative survival probability over several bands and the band-specific rates is also a simple generalization of our earlier result. For a time interval which has been divided into three bands of length T^1, T^2, and T^3, during which the rates are λ^1, λ^2, and λ^3, the log survival probabilities for each band are $-\lambda^1T^1$, $-\lambda^2T^2$, and $-\lambda^3T^3$ respectively. The log of the cumulative survival probability over all three bands is therefore the sum of these, namely

$$-\lambda^1T^1 - \lambda^2T^2 - \lambda^3T^3 = -(\lambda^1T^1 + \lambda^2T^2 + \lambda^3T^3).$$

The quantity $(\lambda^1T^1 + \lambda^2T^2 + \lambda^3T^3)$ is the cumulative failure rate over the whole interval. It follows that the relationship

$$\log(\text{Cumulative survival probability}) = -\text{Cumulative failure rate}$$

still holds when the rate varies from one band to the next.

The use of this relationship to calculate survival probabilities will be demonstrated using the data for the survival of women diagnosed with stage I cancer of the cervix, shown in Chapter 4. The time bands are one year in length and we shall assume that the rate is constant within a time band, but can vary between time bands. Since exact times of failure and loss are not given we shall assume that, on average, each failure contributes 0.5 years to the observation time in the band in which the failure takes place, and similarly for losses. The total observation time during any particular year of follow-up is then approximately

$$
\begin{aligned}
Y &\approx (N - D - L) \times 1 + D \times 0.5 + L \times 0.5 \\
&= N - 0.5D - 0.5L,
\end{aligned}
$$

where N is the number alive at the start of the year, D is the number of deaths, and L is the number of losses during the year. For the first band $N = 110$, $L = 5$, and $D = 5$, so the observation time for the first year is

$$Y^1 \approx (110 - 0.5 \times 5 - 0.5 \times 5) = 105 \text{ woman-years}$$

and the estimated rate is $5/105 = 0.0476$.

For the second band $N = 100$, $L = 7$, and $D = 7$, so the observation time for the second year is

$$Y^2 \approx (100 - 0.5 \times 7 - 0.5 \times 7) = 93 \text{ woman-years}$$

and the estimated rate is $7/93 = 0.0753$.

Exercise 5.7. Estimate the failure rate for stage I subjects during the third year.

The estimated cumulative failure rates for each year of the follow-up are shown in Table 5.2. The column headed 'cumulative survival probability' is obtained using the relationship

$$\text{Cumulative survival probability} = \exp(-\text{Cumulative failure rate}).$$

A life table constructed in this way is sometimes referred to as a *modified life table.*

Exercise 5.8. Calculate the cumulative rate over the last five years only, and hence the probability that a woman survives for ten years *given* that she has survived the first five.

★ 5.6 Rates varying continuously in time

The assumption that the rate parameter is constant over broad bands of time, but changes abruptly from one band to the next, is widely used, but an alternative model, useful when exact times of failure and censoring are known, is to allow the rate parameter to vary from click to click. In Chapter 4 this kind of model led to the Kaplan–Meier estimate of the survival curve; when using rates it leads to the estimate known as the *Aalen–Nelson* estimate.

Fig. 5.4 shows the data that were used to describe the Kaplan–Meier estimate in Chapter 4, but the stepped graph now refers to the cumulative failure rate, not the cumulative survival probability. During the first of these clicks the estimated rate is $0/(50h)$. Similarly for all clicks which contain no failure the estimated rate is zero, so there is no addition to the cumulative rate at any of these points in time. The cumulative rate graph therefore remains horizontal during these clicks. For a click which contains a failure the rate is $1/(Nh)$, where N is the number in the study just before the click. Because this rate operates for a click of length h, the estimate of the cumulative rate increases by

$$\frac{1}{Nh} \times h = \frac{1}{N}.$$

Table 5.2. Modified life table for stage I women

Year	Rate	Cumulative rate	Cumulative survival probability
1	0.0476	0.0476	0.9535
2	0.0753	0.1229	0.8844
3	0.0886	0.2115	0.8094
4	0.0451	0.2566	0.7737
5	0.0000	0.2566	0.7737
6	0.0417	0.2983	0.7421
7	0.0800	0.3783	0.6850
8	0.0000	0.3783	0.6850
9	0.0000	0.3783	0.6850
10	0.0513	0.4296	0.6508

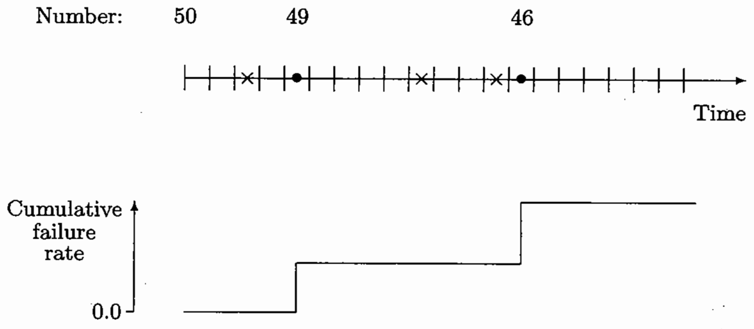

Fig. 5.4. Early follow-up of 50 subjects: the Aalen–Nelson estimate.

Because the click can be thought of as being instantaneous, the cumulative

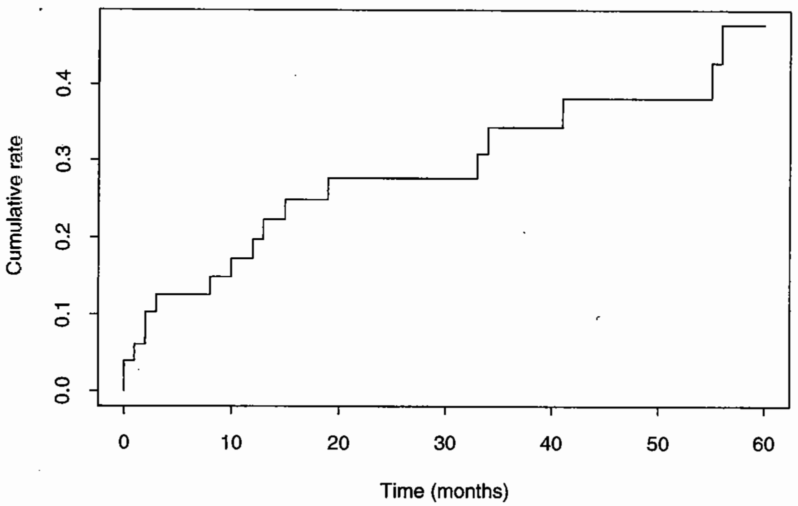

Fig. 5.5. Cumulative rate using the Aalen–Nelson method.

rate jumps by this amount at the moment of occurrence of the failure. In our example, the first jump is of size 1/49; the cumulative rate stays at this value until the click which contains the second failure when it jumps by a further 1/46, and so on.

The cumulative failure rate estimate may also be expressed as a cumulative survival probability, using the now familiar relationship

$$\text{Cumulative survival probability} = \exp(-\text{Cumulative failure rate}).$$

When this is done, the Aalen–Nelson estimate of the relationship of the cumulative survival probability with time looks very similar to the Kaplan–Meier estimate. Both have a stepped shape with steps at the times when failures occur. For most of the follow-up period, the two estimates are very close because of the approximate relationships,

$$\begin{aligned}\log(1 - 1/N) &\approx -1/N \\ \exp(-1/N) &\approx 1 - 1/N\end{aligned}$$

for large N. At the end of the interval N is sometimes small and the two estimates may differ somewhat.

For reasons to be discussed in Chapter 7, it may be best to plot the cumulative failure rate and not the survival probability, even though the former is a little harder to interpret. One fairly clear message from the plot of cumulative failure rate is how the failure rate varies with time. If the failure rate is constant then the cumulative rate will rise linearly with time; if the rate is increasing the cumulative rate will rise non-linearly, showing an increase in gradient with time; if the rate decreases with time the cumulative rate will still rise, but now it will show a decrease in gradient with time.

The Aalen–Nelson plot of the cumulative rate for the melanoma data, introduced in Chapter 4, is shown in Fig. 5.5. This plot shows that the rate is higher during the first 20 months than during the period from 20 to 60 months.

Exercise 5.9. Use the plot in Fig. 5.5 to obtain a rough estimate of the rate during the first 20 months and during the period from 20 to 60 months

Solutions to the exercises

5.1 The total number of subjects observed through one band is

$$7 + 2 + 4 + 2 + 6 + 5 + 10 = 36,$$

and 2 of these end in failure.

5.2 The total observation time for the 30 subjects is 140.1 + 121.8 = 261.9 years. The rate is 30/261.9 = 0.1145 per year, or 114.5 per 10^3 person-years.

5.3 The total observation time is now

$$5 + 5 + 5 + 3.5 + 4.8 + 4.6 + 5 + \ldots + 1.5 = 115.8 \text{ years}.$$

The total number of failures is 14 so the rate is 14/115.8 = 0.1209 per year, or 120.9 per 10^3 person-years.

5.4 The log likelihood at $\lambda = 0.01$ is

$$7 \log(0.01) - 500 \times 0.01 = -37.236.$$

Similarly the log likelihoods at $\lambda = 0.014$ and $\lambda = 0.02$ are -36.881 and -37.384. The log likelihood ratio at $\lambda = 0.01$ is

$$(-37.236) - (-36.881) = -0.3550.$$

Similarly the log likelihood ratio at $\lambda = 0.02$ is -0.5032.

5.5 When the rate is 0.1145 per year, the probability of surviving for 5 years is

$$\exp(-0.11452 \times 5) = 0.564$$

so that the mortality risk is 0.436. The proportion of subjects who failed in this period was, in fact, $14/30 = 0.467$.

5.6 The estimated failure rates for the three bands are 1/13, 0/9, and 1/2 respectively.

5.7 The approximate person-years observation in year 3 is

$$Y^3 \approx 86 - 0.5 \times 7 - 0.5 \times 7 = 79$$

and the estimated rate is $7/79 = 0.0886$ per year.

5.8 The cumulative failure rate over the last five years is 0.173 so that the probability that a woman survives for 10 years given that she has survived the first 5 years is $\exp(-0.173) = 0.841$.

5.9 The gradient of the first part of the cumulative rate curve, from 0 to 20 months, is roughly $0.28/20 = 0.014$ per month, which is the rate over this period (assumed constant). For the second period, from 20 to 60, the gradient is roughly $(0.48 - 0.28)/(60 - 20) = 0.005$ per month, which is the rate over the second period (assumed constant).