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

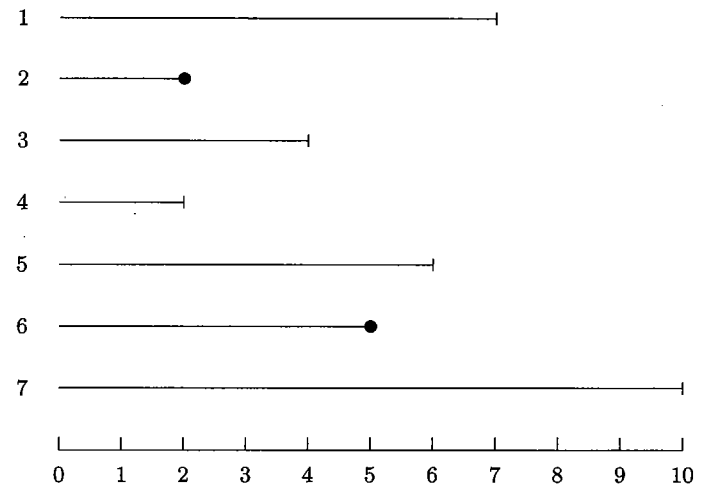

Fig. 5.1. The follow-up experience of 7 subjects.

takes at all time points during the period.

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 $\lambda = 0.02$.

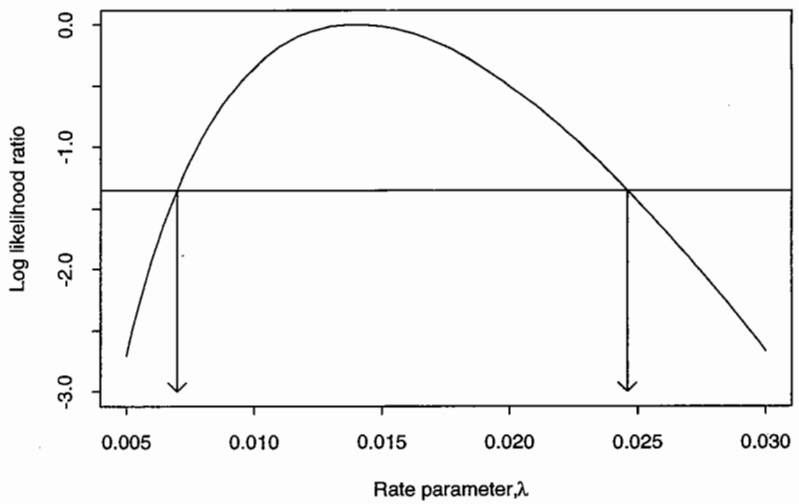

Fig. 5.2. Log likelihood ratio for λ.

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

$$-\lambda N h = -\lambda T.$$

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?

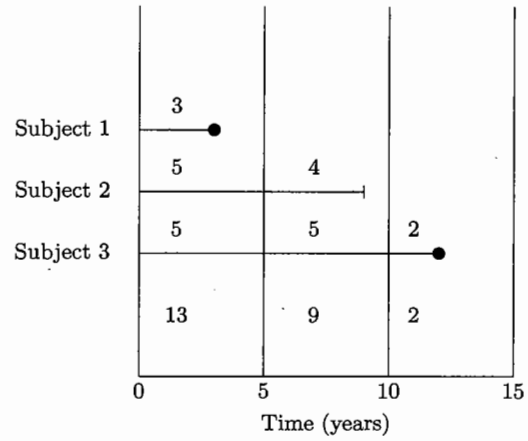

Fig. 5.3. Survival of three subjects across three time bands.

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 $-\lambda^1 T^1$, $-\lambda^2 T^2$, and $-\lambda^3 T^3$ respectively. The log of the cumulative survival probability over all three bands is therefore the sum of these, namely

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

The quantity $(\lambda^1 T^1 + \lambda^2 T^2 + \lambda^3 T^3)$ 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

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

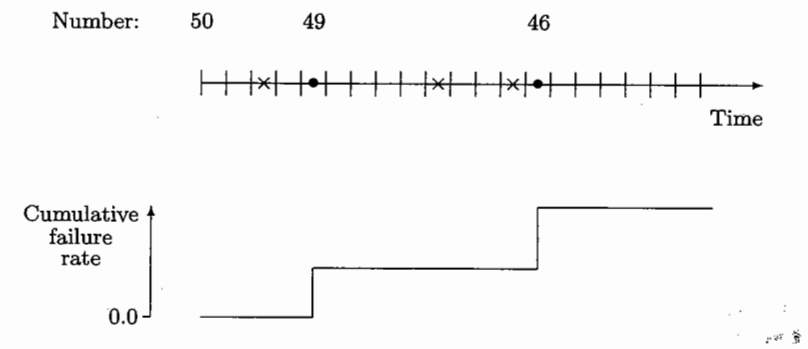

Fig. 5.4. Early follow-up of 50 subjects: the Aalen–Nelson estimate.

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

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

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

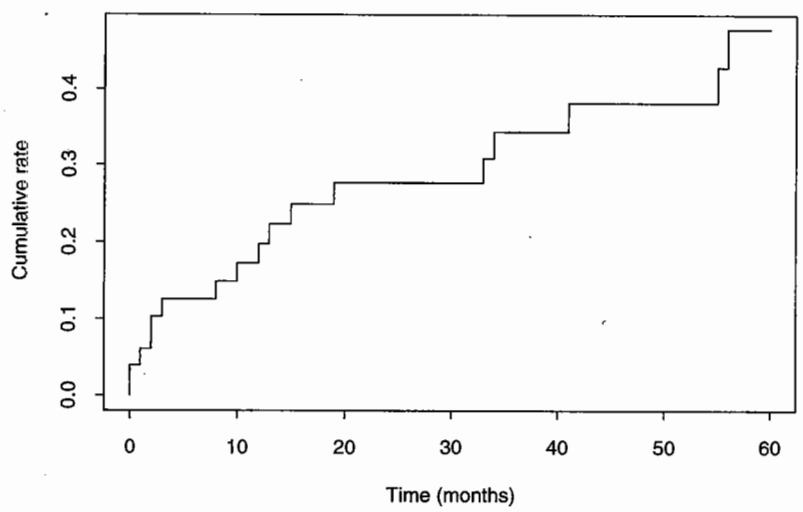

Fig. 5.5. Cumulative rate using the Aalen–Nelson method.

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

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

5 Rates

5.1 The probability rate (hazard rate)

JH is not sure why the authors used the term *probability rate*, when the term *hazard rate*[1], or short-term incidence density, or even just *rate*, or *instantaneous rate*, would have done. The only virtue JH sees for this term is that – unlike the term hazard rate – it is somewhat explanatory: the term does indeed convey, and help you remember, the idea that it is the *probability per unit time*. JH has seen many people struggle to remember and accurately reproduce the definition of the hazard rate. The one item that is not conveyed directly by any of these terms is the *conditional* nature of the probability: it has as its denominator those people, or that person time experience lived by those, who reached the "t" that marks the beginning of the small (infinitesimal) interval.

Another way to think of it is as the limit, as the width of the time band is shrunk to zero, of the incidence density (ID).

Since every realistic and epidemiologically interesting time interval has a non-zero width, and since in any case we usually use the hazard rate as a smooth function of time, the idea of it as an instantaneous rate is merely a mathematical nicety. Indeed, we would immediately multiply this rate into some amount of person time PT (which we can depict as a rectangle with height P persons and width T time units) to get an expected number of events, or for the individual, the conditional probability.[2] The point is that if we were to reverse the process from the expected number of events in a certain PT, the ratio of no. of events to PT would remain the same as we shrunk the width of this time slice, and the corresponding number of events. If it did not, it would imply that the intensity is changing quickly over time, and that a single average intensity (or the corresponding conditional probability) is misleading. See Figure in part I of the 2-part teaching article on going from incidence function to cumulative incidence (a.k.a. 'risk') and back – JH divides the time on each side of a specific t into slices a year, a month, a week and a day wide, and yet the incidence density does not change.

In fact, the force of human mortality is – after a certain age – a monotonically increasing function of attained age (note the conditioning on attained age) but practically speaking, the values of the hazard function at age 32.564 and at 32.565 (or indeed over the age range 32 to 33) are similar enough that we can quite closely approximate this monotonically increasing hazard function (force of mortality) in this age band as a constant, and over a larger age range as piecewise constant within each 1-year age band. If we were concerned with the shape of the hazard function after an attained age or 104, we might want to make the time bands narrower, since the hazard function is 'moving fast' at that age. And at age 32, we might want to make them a bit wider than 1 year: see the value of the q function in the 1-year Canadian lifetables, where q is the conditional failure probability for age bands 1 year wide (h=1 in the terminology of section 5.3)

"The probability rate refers to an individual subject. This is counterintuitive to many epidemiologists."

This is also counterintuitive to JH, who doesn't understand where these authors are coming from on this. An incidence density is certainly not about an individual person[3]. How are we to think of a failure rate of 8 ruptures per 10000-pipe-kilometer-years of operating pipeline of a water distribution system?

The authors however do well to ask us to distinguish between the definition of the *parameter*, and an *estimate* (or estimator) of the value of this parameter in a particular context (e.g. the rupture rate when the temperature is in the vicinity of -20C.)

Mathematically, then, here are a few definitions of what they call the probability rate, or simply the instantaneous rate, at time t. Since it is a parameter, we will, as they do, give it the Greek letter lambda, λ. With P the number of persons at risk at t, or more realistically, the average number of persons at risk over the entire interval $(t, t+\delta t)$,

$$\lambda(t) = \lim_{\delta t \to 0} \frac{\text{Expected no. of events}}{P \times \delta t}$$

[1] The Website `jeff560.tripod.com/h.html` "Earliest Known Uses of Some of the Words of Mathematics" tells us: HAZARD RATE came into use in statistics in the 1960s as a general term for what is called the force of mortality in demography and the intensity function in extreme value theory. David (2001) finds "hazard rate" in R. E. Barlow; A. W. Marshall & F. Proschan "Properties of Probability Distributions with Monotone Hazard Rate," Annals of Mathematical Statistics, 34, (1963), 375-389. A JSTOR search found "death-hazard rate" in D. J. Davis "An Analysis of Some Failure Data," Journal of the American Statistical Association, 47, (1952), 113-150.

[2] Freedman, in his nice article, Survival Analysis: A Primer" in the American Statistician in May 2008 (see resources for survival for course `EPI634`) puts it nicely: "The intuition behind the formula is that $h(t)dt$ represents the conditional probability of failing in the interval $(t, t+dt)$, given survival until time t."

[3] There is, to some epidemiologists, a difference between the value for the collective, and the value for the individual. British medical statistician William Farr (1807-1883) and McGill epidemiologist Olli Miettinen (1936-) both take what we now think of as the '*cumulative incidence*' proportion to refer to an empirical or theoretical value for a *collective*, whereas when an *individual* uses that value as his/her own probability, it should be called a *risk*.

One can re-write this as

$$\lambda(t) = \lim_{\delta t \to 0} \frac{\text{Expected no. of events}}{P} \div \delta t$$

so that the Expected no. of events/$Person$ looks somewhat like a probability. This probability, when divided by δt becomes the (conditional) failure probability *per unit time* that the authors use as their definition.

One will also see in survival analysis textbooks the definition of $\lambda(t)$ or $h(t)$ as

$$h(t) = \lambda(t) = f(t)/S(t),$$

where $S(t)$ is the 'survival' function, i.e., $1 - F(t)$, and $f(t)$ the probability density function, of the 'time to event' random variable. This is no different from the definition above, since we can write it as

$$h(t) = \lambda(t) = \frac{f(t)\delta t}{S(t)} \div \delta t.$$

S(t) is the proportion of persons who are at risk (event-free) at time t, and $f(t)\delta t$ is the (unconditional) fraction of events that occur within the interval $(t, t + \delta t)$, so $\frac{f(t)\delta t}{S(t)}$ is itself a (conditional) fraction of a fraction.

Moreover, we can rewrite the definition as

$$h(t)dt = \lambda(t)dt = \frac{-dS(t)}{S(t)}$$

and integrate both sides over the interval $(0, T)$ to get

$$\int_o^T h(t)dt = \int_o^T \lambda(t)dt = \int_o^T \frac{-dS(t)}{S(t)} = -\log S(T).$$

Then, exponentiating both sides, we get the **fundamental relationship between the incidence density function (alias hazard function ($h(t)$, or the maybe more familiar term 'failure rate function', $\lambda(t)$) and the complement of cumulative incidence (CI)**[4] , namely

$$1 - CI_{0\to T} = S(T) = \exp\left[-\int_o^T h(t)dt\right] = \exp\left[-\int_o^T \lambda(t)dt\right].$$

Notice also the (welcomed) use throughout the book of λ as an event *rate*, and not – as some books use it – as the expected *number* of events, i.e. as the mean parameter of a Poisson distribution. JH has tried to *be consistent in using the Greek letter μ for the expected number of events*, since after all it is the mean or expected value of the random variable, and since it is important to keep the distinction between the numerator and denominator of an event rate parameter.

5.2 Estimating *the* probability rate parameter

Notice the use of the (underlined) word *the*, i.e., that the parameter value is assumed constant in the follow-up period of interest.

5.3 The likelihood for a rate parameter

You might find it strange that the authors don't go directly to the representation of the observed rate as an observed Poisson numerator divided by a known PT denominator. I think they did this to emphasize the idea of subdividing the PT into person-clicks.

It is interesting that in 1907 Gosset (of 'Student-t' fame) derived the Poisson distribution 'from scratch' using this same conceptual subdivision of a plate (or field in a microscope) into a large number of small squares, small enough that only one yeast cell would fit in it (C&H in section 4.4 write of time bands so narrow that "each failure occupies a band by itself").[5] If the mean number

[4] Ways to 'see' this relationship in heuristic terms are described in part II of the draft teaching article. JH has been searching a long time for who might have been the first to derive this relationship: as JH notes in the article, Chiang says that the equation

> has been known to students of the lifetable for more than two hundred years. Unfortunately, it has not received much attention from investigators in statistics, although various forms of this equation have appeared in diverse areas of research.

As of now (October 2012), JH believes that Chiang was probably referring to a paper by Daniel Bernoulli published in 1766, where he calculates the gain in life expectancy after elimination of this cause of death (smallpox). His solution to that more complex problem involves the solution of the same differential equation we discuss above. See website for more details.

[5] JH has put this very readable 1907 article "*On the Error of Counting with a Haemacytometer*" under the resources for rates in course `EPI634`

of cells per plate was μ and the area of the plate was A, or $N = A/a$ small squares of area a each, then the probability π that a small square contains a square is $\pi = \mu/N$. The probability that the total area A will contain y yeast cells is then

$$Pr(y \text{ occupied cells}) = {}^{N}C_{y}\ \pi^{y}(1-\pi)^{N-y}.$$

Gosset used Stirling's approximation, and the definition of $e^x = \exp[x]$ as a limit, to go from this binomial probability to the Poisson probability $\exp[-\mu]\ \mu^y/y!$

If we worked with μ directly, then (ignoring the factorial, which doesn't involve this parameter), the likelihood based on an observed count of D is

$$\exp[-\mu]\ \mu^{D}.$$

Substituting $\mu = \lambda Y$, where Y is C&H's notation for amount of person-Years (what we call the denominator) gives

$$\exp[-\lambda Y]\ (\lambda Y)^{D},$$

or, ignoring items that do not involve λ, as

$$\exp[-\lambda Y]\ (\lambda)^{D},$$

so that the log-likelihood is indeed

$$-\lambda Y + D \log\ (\lambda),$$

It is interesting to go back to the derivation (section 81, pp. 205-206) by Poisson in his 1837 book Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilitiés (Paris, France: Bachelier, 1837). You can read the original via the Wikipedia link `http://en.wikipedia.org/wiki/Poisson_distribution`. Poisson also starts with the binomial, and goes to "le cas où l'une des deux chances p et q est très petite."

5.3.1 Example: Likelihood for parameter of exponentially distributed random variable, with interval censoring.

The Uganda and Kenya 'circumcision in the prevention of HIV' studies are examples of interval-censored (as well as the usual right-censored) data, since one cannot know exactly when a person became HIV+, only that it occurred in the interval between the last negative test and the first positive one.

Before setting up the likelihood for such data, let us consider a simple statistical model for the data, and let us focus for now on the placebo group. *We will assume that the sero-conversion rate λ is constant over the 2 years,* i.e., that $\lambda(t) = \lambda$ over that interval. Up until now, we treated the number of events in the 'aggregated-across-subjects' person time as a Poisson random variable. *Another way to look at this is to consider the inter-event times, (or the time-to-event times) and their distribution.* We know from BIOS601 that if the event rate is λ, and there is always one unit at risk, then the inter-event times have an exponential distribution with mean $1/\lambda$. Thus, we can say that the 'time-to-event' for each subject is a realization of an exponential random variable with mean or expected value $1/\lambda$. If we call this r.v. 'T', then

$$T \sim \exp(\mu_T = 1/\lambda),$$

$$S_T(t) = \exp[-\lambda t],$$

$$F_T(t) = 1 - S_T(t) = 1 - \exp[-\lambda t],$$

$$f_T(t) = F'_T(t) = \lambda \exp[-\lambda t] = (1/\mu_T) \exp[-(1/\mu_T)t].$$

In the control group in the Uganda trial, 2319 initially HIV- men were tested at the 6-month, or 0.5year follow-up, and 19 of them were found to be HIV+, and the remaining 2300 were found to be HIV-.

The likelihood, based just on this first follow-up test is therefore the probability (as a function of the seroconversion rate λ) of observing this pattern of results. First we write it as a product of 2319 probabilities:

$$Likelihood = \prod_{i=1}^{i=2319} Pr[obs'd\ outcome\ for\ subject\ i] = \prod_{i=1}^{i=19} Pr_i \prod_{i=20}^{i=2319} Pr_i$$

With T denoting the r.v. 'time to HIV+', each Pr_i in the second product is of the form $Pr[T > 0.5 \mid \lambda] = \exp[-0.5\lambda]$, while each Pr_i in the first product is of the form $Pr[T < 0.5 \mid \lambda] = 1 - \exp[-0.5\lambda]$. The likelihood based on this first test can thus be simplified to

$$L_{1st\ test} = \exp[-2300 \times 0.5\lambda]\ \times\ (1 - \exp[-0.5\lambda])^{19}$$

Some 2229 of those HIV- at 6-months were tested at the 12-month, or 1-year follow-up, and 14 of them were found to be HIV+, and the remaining 2215 were found to be HIV-. Thus the likelihood based on this second test can thus be simplified to

$$L_{2nd\ test} = \exp[-2215 \times 0.5\lambda]\ \times\ (1 - \exp[-0.5\lambda])^{14}$$

Notice that with this exponential distribution, the fact that these 2229 had got throught the first interval HIV-free has nothing to do with their (now conditional) probabilities for the next 6 months. Technically, we call this the "memoryless" property of the exponential distribution.[6] Thus, $Pr[T > t \mid T > t_{given} = Pr[T > t - t_{given}]$, and so, whereas we would normally have to use the *conditional* probability $\{F(1.0) - F(0.6)\}/S(0.5)$, here we can use the unconditional probability of escaping infection for 6 months. In effect, we can 'reset the clock to zero at T=0.5,' and imagine it was just like back at $T = 0$.

Some 980 of those HIV- at 12-months were tested at the 24-month, or 2-year follow-up, and 12 of them were found to be HIV+, and the remaining 968 were found to be HIV-. The likelihood based on this third test can thus be simplified to

$$L_{3rd\ test} = \exp[-968 \times 1.0\lambda] \ \times \ (1 - \exp[-1.0\lambda])^{12}$$

Thus the likelihood based on all three tests is

$$L_{all\ 3\ tests} = L_{1st\ test} \ \times L_{2nd\ test} \ \times \ L_{3rd\ test}$$

ie

$$L = \exp[-(2300 \times 0.5 + 2215 \times 0.5 + 968 \times 1.0)\lambda]$$
$$\times$$
$$(1 - \exp[-0.5\lambda])^{12} \ \times \ (1 - \exp[-0.5\lambda])^{14} \ \times \ (1 - \exp[-1.0\lambda])^{12}$$

Supplementary Exercise 5.1.

1. Maximize L with respect to λ.
2. What would happen to L, and to the ease of estimation, if subjects were tested more frequently, e.g. every month, every week, every day?
3. Superimpose the smooth cumulative incidence [also called the 'risk'] curve, CI(t), derived from the exponential model for the 'time to HIV infection' (or, equivalently, the constant-over-time infection rate model) on the step-function curve in the article. If you were a co-author, Which of the two curves would you would suggest be presented?

[6] In industrial life-testing, this property is referred to as the 'used is the same as new' property. In failure time distributions where the failure is a function of age or duration of use (e.g. a computer or hard disk), the hazard is — maybe after a certain run-in period – an increasing function of its age or accumulated hours of work, and so the testers say 'older is worse (less 'reliable') than newer;' initially, before those units doomed to early failure have been weeded out, it may be that 'newer is worse than older.' Sadly, most human hazards, other than being struck by a meteor, are from internal sources to do with our own bodies, and so while the hazard function or force of mortality decreases until about age 8 – see Canada lifetables – it is monotonically increasing thereafter.

Supplementary Exercise 5.2.

Refer to the data, kindly supplied by the author, on the 78 Icelandic trios studied by Kong et al. These, along with the original Nature article 'Rate of de novo mutations and the importance of father's age to disease risk' can be found in the Resources link for C&H Ch05 [Rates: N-A estimate].

1. Assume that the *de novo* mutation rate (λ) is independent of (constant over) a man's age (life),

   $$\text{Mutation Rate at Age } a = \lambda, \ \forall a,$$

   and that the mutations found in his children are all transmitted from him, and that none are inherited from the children's mother.

   Use the rate estimator you derived from the '2 data-points and a (Poisson) model' exercise earlier in the term to estimate the mutation rate from the data on the 78 trios. This empirical rate involves very simple grade 6 arithmetic, using the 'sufficient' statistics. see if you get the same parameter estimate from a 'canned' regression program that uses the individual-subject data.

2. Derive ML estimators for the two parameters λ_0 and β in the age-dependent mutation rate models:

   $$\text{Mutation Rate at Age } a = \lambda_0 + \beta \times a, \qquad (\textit{additive rate model})$$

   and

   $$\text{Mutation Rate at Age } a = \lambda_0 \times \exp[\beta \times a], \qquad (\textit{multiplicative model}).$$

   For the additive model, check that applying your estimator directly to the data (i.e., by coding the formulae in `R`) yields the same parameter estimates as you would get from a 'canned' regression routine.

5.4 Cum. survival probability as fn. of rate parameter

We saw this earlier as $S(T) = \exp[-\int_0^T h(t)dt]$, or cumulative incidence as $CI_{0\to T} = 1 - S(T) = 1 - \exp[-\int_0^T h(t)dt]$.

We also came up with a 'heuristic' ("a usually speculative formulation serving as a guide in the investigation or solution of a problem") whereby **the integral $\int_0^T h(t)dt$ can be seen as the expected number of events, μ, if there was always one unit (person) at risk for the period 0 to T.** Thus if an event (failure) occurred at any point in this interval, the failed unit is immediately replaced by another of the same profile: e.g., if $h(t)$ referred to computers, we would replace a computer that failed at time t_1 by another of the same age, and if this failed before T, at time t_2 say, we would in turn replace it by another of age t_2, and so on until we got to T. So by the end, we would have observed the 1-unit system for a total of T units of time, and we might have observed $0, 1, 2, \ldots$ failures (and had to make this many *replacements*), in order to have the system in continuous operation for this duration. The expected number of failures in that period would be the integral of (the area under) the $h(t)$ curve. We saw earlier that the Poisson distribution has the 'closed under addition' property; in this application, we can think of the total number of events in $(0, T)$ as (the limit of) a sum of more and more Poisson random variables, representing the numbers of events in smaller and smaller intervals $(t, t + dt)$, with expected numbers of events $h(t)dt$. In the limit, this sum of small expectations is nothing more than the overall expected number of events,

$$\mu = \int_0^T h(t)dt$$

The observed sum (the total number of replacements) is thus the realization of a single Poisson random variable with mean μ, and so the probability that the initial unit will 'survive' the entire interval is just the probability that there will be no event in the entire period, i.e.,

$$S(T) = Pr(Poisson.RV[\mu] = 0) = \exp[-\mu] = \exp[-\text{integral of } h(t)].$$

The other concept that is reinforced by this heuristic, and the computer example, is that the computer-days are interchangeable. Imagine we had a large bank of computers all of the same vintage: we could imagine having a different one of these computers be the one that ran the system (was 'on duty') for the day, and we could even draw lots for which computer is the one on duty at any time. Assuming that the 'on duty' computer didn't age any faster than the ones that were 'off duty' that day, we can now see that the probability that a *specific* computer would fail before time T is the same as the probability that a *sequence of computer-days – or computer-hours, or computer-minutes* (each one contributed by a possibly different computer) would contain at least one failure. This *interchangeability* of (impersonal, indistinguishable, unnamed) units of the same age, i.e., with the same $h(t)$, is central to the concept of 'person-clicks' that C&H use.. it is not the particular person that matters to the contribution, but the person's *profile* – his/her $h(t)$ value.

If the rate is a constant over the period $(0, T)$, so that the integral is $\mu = \lambda \times T = \lambda T$, then we get the simple expression for the (cumulative) survival probability given at the top of page 46, namely $S(T) = exp[-\lambda T]$.

This section also discusses the simple approximation to $exp[-\mu]$ when μ is small, namely $1 - \mu$. In this situation, the cumulative risk (in fact, the word *cumulative* is redundant!) can thus be approximated by

$$\text{Risk} = \text{Cumulative Incidence } \approx 1 - \mu = 1 - \lambda T \quad [\mu \text{ small}].$$

Whether or not the integral μ is small, if λ is constant over $(0, T)$, then – apart from random variations –

$$\log\{S(t)\} = \log\{\exp[-\lambda t]\} = -\lambda t,$$

so that

$$\text{the plot of } -\log\{S(t)\} \textit{ versus } t \text{ should be linear in } t, \text{ with slope } \lambda.$$

5.5 Rates that vary with time

JH's comments in section 5.4 discussed both piecewise-linear and (in the limit) general smooth form(s) for $h(t)$ or $\lambda(t)$, and so there is little to add for this section, other than to make one remark about their use of the term "*cumulative failure rate*." JH finds this term too close to "cumulative incidence", which is a proportion. C%H's "cumulative failure rate" is in fact the integral we discussed above, and so has as its dimension or units the expected number of events in the period $(0, T)$ if one unit were always operating, i.e., 'at risk.' He would prefer that you use the more common term "*integrated hazard*" often denoted by an upper case letter,

$$H(T) = \int_0^T h(t)dt \quad or \quad \Lambda(T) = \int_0^T \lambda(t)dt.$$

C&H tell us that "it follows that the relationship

$$\log[S(t)] = -Cum.\ failure\ rate \ \ \{\ \log[S(t)] = -H(t) \ \ in\ our\ notation\ \}$$

still holds when the rate varies from one band to the next... and will be used to calculate $S(t)$." We have already used the exponentiated version of this

to calculate $S(t)$. But this relationship in the log scale is also used to check whether an assumed form or model for $h(t)$ fits with the observed data: it is more difficult to judge fit on the S scale, where $S(t)$ is likely to be quite curvilinear, than on the H scale, where $H(t)$ may have a simpler form, such as piecewise linear.

Supplementary Exercise 5.3.

1. For the Uganda HIV data, assume a different λ for each of the 3 intervals, and estimate each one separately. Do the data provide evidence against this assumption? Answer by maximizing L under the larger (3 possibly different λs) and smaller (all three λs are the same) models, and computing the likelihood ratio.

2. Even if you do not find evidence against a constant-over-time-bands λ model, nevertheless calculate and plot the (piecewise-smooth) cumulative incidence CI(t), derived from the '3-λ' model, superimpose it on the CI(t) curve fitted under the simpler '1-λ' model.

5.6 Rates varying continuously in time: Kaplan-Meier (K-M) and Nelson-Aalen (N-A) estimators

"*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.*"

This is a very nice way of putting it. First, it says that the Kaplan-Meier curve is a limiting case of a probability-based lifetable, with the time bands made narrower and narrower. In the limit (and the Kaplan-Meier table is sometimes referred to as the 'product-limit' table) one need only be concerned with products of continuation probabilities from the event-containing intervals. It also explains why the Kaplan-Meier curve is called 'non-parametric': by making the bands narrower and narrower, the curve follows the data exactly.

The Kaplan-Meier estimate can be seen as a product of *empirical* continuation *probabilities*, each one governed by the *binomial* model. We formally acknowledge this when we use Greenwood's formula for the SE of $\widehat{S(t)}$.

The Nelson-Aalen estimate can be seen as a product of model-based continuation *probabilities*, with each estimated probability calculated from the theoretical relation between the (in this case shortterm incidence or) hazard rate and cumulative incidence, viz. $S_{t\to t+dt} = 1 - CI_{t\to t+dt} = \exp[-\int_t^{t+dt} h(u)du$

If an interval $t, t+dt)$ involves n persons at risk, and d events (deaths), then the person time is ndt and so the estimate of the incidence is $\frac{d}{n\times dt}$. each one governed by the *binomial* model. If d is zero, then the estimate of the incidence is zero. Thus, the empirical hazard function is a square-wave function,

$$\widehat{h(t)} = \begin{cases} 0 & \text{if } (t, t+dt) \text{ contains } d = 0 \text{ events,} \\ \frac{d}{n\times dt} & \text{if } (t, t+dt) \text{ contains } d > 0 \text{ events.} \end{cases}$$

Thus,

$$\widehat{h(t)dt} = \begin{cases} 0 & \text{if } (t, t+dt) \text{ contains } d = 0 \text{ events,} \\ \frac{d}{n} & \text{if } (t, t+dt) \text{ contains } d > 0 \text{ events.} \end{cases}$$

Thus

$$\int_0^T \widehat{h(t)dt} = \sum \frac{d}{n},$$

with the summation over those event-containing narrow bands where $t < T$. The persons at risk in these event-containing bands are called *risksets*.

The `EPIB634` site has R code that divides the JUPITER follow-up time into 1-year, then 1-month, then 1-week, then 1-day bands. The resulting h(t) function becomes more and more erratic, but in doing so – just like the K-M curve – it conforms exactly to the data.

Just as the K-M curve is based on a product of *binomial*-based probability estimates, the N-A curve can be seen as an integral (the limit of a sum) of *Poison*-based rate (hazard) estimates: provided that each n is large, the 'd' that forms the numerator of the empirical elemental area can be seen as a realization of a Poisson random variable. Its estimated variance can therefore be estimated as d, and the variance of $\frac{d}{n}$ as $\frac{d}{n^2}$. Thus,

$$\widehat{Var}\left[\int_0^T \widehat{h(t)dt}\right] = \sum \frac{d}{n^2}.$$

For the numerators in this variance expression, some textbooks use binomial-based variances of $n \times \frac{d}{n} \times \frac{n-d}{n}$ instead of the Poisson-based variances of d. If each $n-d$ is large, as it is in the JUPITER study, then the difference between the two formulations is miniscule.

Most software packages plot the N-A curve as a step-function, just as they do the K-M curve. The conf. intervals are first calculated for the estimated integral, and then for $\widehat{S(t)}$.

Supplementary Exercise 5.4.

1. Calculate the Nelson-Aalen and Kaplan-Meier curves, and the SE's, for the placebo arms of the Uganda and Kenya circumcision trials, and the JUPITER trial.

5.7 'Lifetime' (and Portion-of-Lifetime) Risks

Supplementary Exercise 5.5.

A recent bios601 seminar addressed the risk of appendicitis in twins. It drew on the self-reported (in a 1980 survey) experience of Australian twins. The original paper can be found at `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1683858/` and the data and documentation at `http://genepi.qimr.edu.au/staff/davidD/Appendix/`.

1. Using the supplied `R` code, or otherwise, and using the appendectomy data from all respondents, calculate the age-specific hazard rates, i.e., the incidence density as a function of age. Use all respondents.

2. What is (i) the average age ($\overline{a}$) of the respondents at the time of the survey? (ii) the average (and IQR) age and calendar year at/in which the appendectomies were performed? Using the age-specific (hazard) rates to calculate the cumulative incidence of appendectomy to ages 25, 30, 35 and 40, and to age $\overline{a}$. Compute the observed overall proportion in the dataset who have had an appendectomy, and comment on how well it agrees with the 5 fitted values.

3. Repeat the cumulative incidence estimation in 2., but with a suitably smoothed hazard function.

4. Repeat the cumulative incidence estimation in 2., but using the Kaplan-Meier estimator.

5. Refer to the Norwegian article "Incidence of Acute Nonperforated and Perforated Appendicitis: Age-specific and Sex-specific Analysis" by Körner et al. in World J. Surg. 21, 313-317, 1997. Use the data in Table 1 to calculate the (sex-specific) cumulative incidence of laparatomy for suspected appendicitis to ages 25, 30, 35 and 40. State any assumptions you make. Compare the estimates with the ones based on the Australian data.

6. Quickly examine the articles, provided under Resources, from other places and times – and in some instances using slightly different 'events'. Make a few quick back-of-the-envelope calculations of the risks (lifetime, and to ages 25, 30, 35 and 40) they imply. Then comment on where the Australian-derived estimates fit in relation to all of the other estimates, and whether the discrepancy can easily be explained in terms of differences in 'event-definition' or different 'persons, or places or times'. Could you imagine some selectivity in who responded to the survey?