4 Consecutive follow-up intervals

4.1 A sequence of binary models

The lifetable as a sequence of Bernoulli models: Efron (1977) was one of the early authors to point out that the likelihood contribution of a subject, followed for t units of time, is equivalent to the likelihood for a sequence of a large number, $n = t/\Delta$, of Bernoulli trials, with time-dependent probabilities of failure. For a trial that corresponds to the small interval $(t, t+\Delta)$, the failure probability can be well approximated by $p = h(t)\Delta$, where $h(t)$ is called the hazard function (see later). The sequence ends with the n^{th} trial, at the time of the event of interest or when follow-up was otherwise terminated. In a subsequent article Efron (1988) focused on discretizations of the t-axis and on using logistic regression to fit various smooth-in-t hazard and survival functions in the one-sample situation, where the usual non-parametric alternative is the Kaplan-Meier estimator of survival rate.

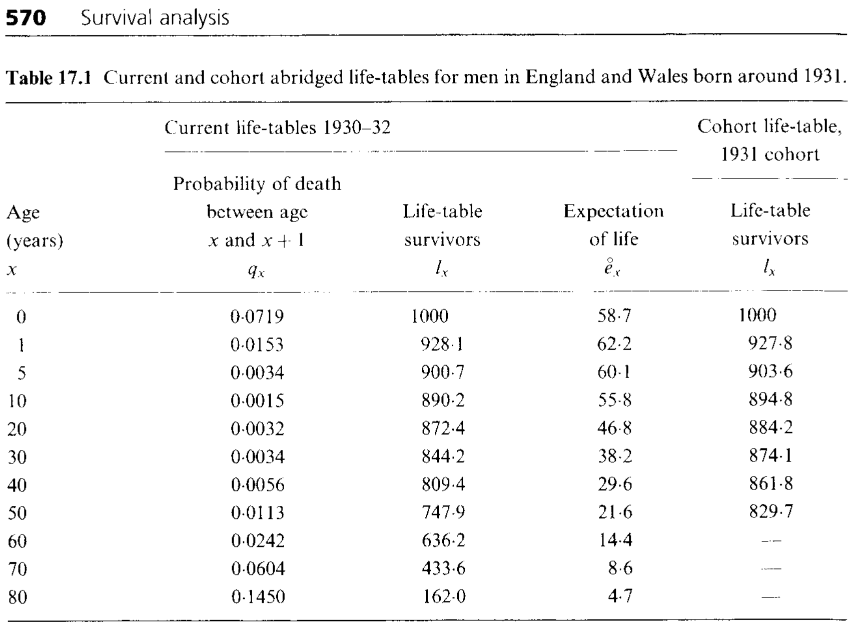
The probabilities of surviving one, two, and three years without failing are called the *cumulative survival probabilities* for the cohort: JH continues to argue that the word cumulative is misleading. The complement of the (unconditional) survival probability is the *cumulative incidence.* It is an increasing function. Would we call a declining fraction, obtained as a product of more and more fractions, a *cumulative* fraction?

4.2 Estimating the conditional probabilities of failure

The subjects who contribute to the estimation of the conditional probabilities do not have to have been followed from the beginning. One can splice together estimates based on separate samples. This is what is done to create current lifetables. And in any case, when (a subset of) those who "survive" a specific time band are used again in the next band, the estimates are treated as independent of each other – just as if they were from different persons. In current lifetables, they are different persons!

Table 17.1 in p. 570 of the Survival Analysis chapter (17) of the 4th edition of Statistical Methods in Medical Research by Armitage, Berry & Matthews, illustrates the difference between 'current' (aka 'period') and 'cohort' lifetables.

The entire 'current' lifetable is calculated, as a product of conditional probabilities, using the *observed* age-specific mortality rates in England and Wales in 1930-1932. In this sense it is fictitious, since those who computed the table in the 1930's didn't know for sure that the world would even exist in 2010, when those remaining from the fictional 1000 who started out at age 0 would reach their 80th birthday. And even if they did, they could not have anticipated exactly what force of mortality these 80-year olds would face in 2010, even though they might have foreseen that mortality rates would improve over time. The force of mortality these 80-year olds would face in 2010 is a good deal lower than the force of mortality the 80-year olds actually faces in 1930-32. For example, the death rate in the male 75-79 age category in Denmark was 9.4/100MY in 1930-34 and 4.2/100MY in 2000-04.



570 Survival analysis

Table 17.1 Current and cohort abridged life-tables for men in England and Wales born around 1931.

	Current life-tables 1930–32			Cohort life-table, 1931 cohort
Age (years) x	Probability of death between age x and $x+1$ q_x	Life-table survivors l_x	Expectation of life $\mathring{e}_x$	Life-table survivors l_x
0	0·0719	1000	58·7	1000
1	0·0153	928·1	62·2	927·8
5	0·0034	900·7	60·1	903·6
10	0·0015	890·2	55·8	894·8
20	0·0032	872·4	46·8	884·2
30	0·0034	844·2	38·2	874·1
40	0·0056	809·4	29·6	861·8
50	0·0113	747·9	21·6	829·7
60	0·0242	636·2	14·4	—
70	0·0604	433·6	8·6	—
80	0·1450	162·0	4·7	—

"The cohort life-table describes the actual survival experience of a group, a 'cohort' of individuals born at about the same time. Those born in 1900, for instance, are subject during their first year to the mortality under 1 year of age prevailing in 1900-1; if they survive to 10 years of age they are subject to the mortality at that age in 1910-11; and so on. Cohort life-tables summarize the mortality at different ages at the times when the cohort would have been at these ages. The right-hand side of Table 17.1 summarizes the l_x column from the cohort life-table for men in England and Wales born in the 5 years centred around 1931. As would be expected. the values of l_1 in the two life-tables are very similar, being dependent on infant

> mortality in about the same calendar years At higher ages the values of l are greater for the cohort table because this is based on mortality rates at the higher ages which were experienced since 1932."

For a further illustration of the difference between 'current' and 'cohort' life tables, see the Bridge of Life applets (accessible via link at bottom left of JH's homepage). In particular, see the contrast between France, 1895 (current) and France, 1895-2004 (cohort).

This exercise makes it clear that, in the analysis of such studies, the ***basic atom of data*** *is not the subject, but the observation of* ***one subject through one time band****.* [last para of section 4.2]

This is a very important statement, and this 'outlook' or 'attitude' is key to a full understanding of rates, and or person-time. It says that one's 'timeline' is divisible. Think of the experience as an infinite sequence of Bernoulli trials that is terminated by the event, or when observation is terminated (i.e., before the event could occur).

It also allows the experience to be further sub-divided into 'exposed' person time bands and 'unexposed' person time bands: c.f. of the 'clicks' of time a driver spends on the cell-phone and off-the-cell-phone.

In the example, the event of interest is a one-time event, and so, unlike the cat with nine lives, once the event occurs, it terminates the observation: one is no longer 'at risk.' But one can also think of events, such as repeated events such as accidents, or sickness episodes, experienced by the same person.

4.3 A cohort life table

These [survival] plots are useful for studying whether the probability of failure is changing with follow-up time, and for calculating survival probabilities for different periods of time. In fact, it is not that easy to check if the probability of failure is changing from survival curves. The probability of failure the authors write of is a *conditional*, i.e. time-specific, probability, and so the hazard function, which uses as a denominator the numbers of persons *at risk at that time*, makes it easier to monitor this probability.

4.4 The use of exact times of failure and censoring

"[...] *choosing the bands so short that each failure occupies a band by itself.*" This is the same assumption that allows us to derive the Poisson distribution as a limiting case of the Binomial distribution, and the link between the Poisson distribution and the exponential distribution of inter-event times.

"*This method of estimating the cumulative survival probabilities is called the Kaplan-Meier method*" It is also called the product-limit method, since it is derived by slicing time into smaller and smaller bands, and not having to be materially concerned about where within the band an observation becomes censored. In the JUPITER trial example JH is using in the EPIB-634 course, the follow-up ranges from just over a year to almost 5 years, or approximately 400 to 1600 days. The 200+ events in the placebo arm, and the 100+ in the treatment arm, are distributed over these 1600 days. If we use one day as the width of each band, and estimate $S(1000)$, the 1000-day "event-free survival" then this estimate is a product of 1000 estimated conditional probabilities, many of them estimated at unity. *So the changes in the product take place only at the days in which there were events.* See also the COMPARE trial.

The persons at risk just before the event on a particular day (including the person(s) who did suffer the event that day) are called the *riskset*. They are the *candidates* for the event.

Supplementary Exercise 4.1 Consider again the tumbler longevity data that we saw in an earlier exercise. The smallest unit of time ('granularity') is 1 week. Even though some observations ($< 10\%$) are right-censored, Table 1 in the paper lists the data in a form that allows direct calculation of an empirical complement-of-the-cdf by 'coarse-products' rather than exact product-limits. Graphically compare the results obtained with this (non-parametric) estimator of the 'the survival' function with the results obtained with the (parametric) gamma model fitted by the authors. Compare also the mean longevity estimated by calculating the area under the non-parametric survival curve with that obtained from the values of the 2 fitted parameters of the author's model.

4.4.1 $\widehat{S(t)}_{KM}$: a Non-parametric Maximum-Likelihood Estimator (NPMLE) of $S(t)$

As is rigorously justified in their 1958 paper, the Kaplan-Meier estimator is a non-parametric ML estimator within the class of all possible $S(t)$ functions.

Supplementary Exercise 4.2 Take a small survival dataset with just 3 observations, 1 censored and 2 not, such as the 3 values 5, 7+ and 10. Show that

$\widehat{S(t)}_{KM}$	Interval	Point (t)	Prob. Mass at Point
1	$t < 5$		
		$t = 5$	1/3
2/3	$5 \leq t < 10$		
		$t = 10$	2/3
0	$t \geq 10.$		

maximizes the Likelihood, ie the probability of the observed data as a function of $S(t)$, i.e., that no other $\widehat{S(t)}$ can yiled a larger likelihood.

4.4.2 $\widehat{S(t)}_{KM}$ as a '<u>self-consistent</u>' and as a <u>Distribute</u> <u>mass</u> <u>to</u> <u>the</u> <u>right</u>' estimator of $S(t)$

The K-M estimator, based on n observations $T_1, \ldots, T_n$, some censored, some not, can also be seen as obeying the self-consistent estimating equation:

$$S(t) = \frac{1}{n}\left\{\sum_{all} I[T_i > t] + \sum_{censored\ <\ t} \frac{S(t)}{S(T_i)}\right\}$$

Observations known to exceed t [even if censored after t]are counted as survivors (1's) while observations for which we don't know if they will exceed t are counted as fractions or probabilities: those which are already close to reaching t are given higher chances of eventually exceeding it, those which are further to the left of t are given lower chances of doing so: the chance of eventually exceeding t, given that one has already reached a value $T < t$, is $S(t)/S(T)$.

The K-M estimator can also be seen as a **distribute to the right** procedure: Initially, each of the n observations is given a mass of $1/n$. Then, the mass given to the leftmost censored observation is redistributed (equally) to all observations to the right of it, and that leftmost observation is removed. The process is repeated until all censored observations are removed, and all of their mass has been redistributed.[16]. The procedure will remind some of the EM algorithm.

Supplementary Exercise 4.3 Take a simple survival dataset with just 5 observations, 2 censored and 3 not, such as the 5 values 2, 5+, 6, 7+ and 9. Derive the K-M estimate of $S(t)$. Illustrate the 'self-consistency' of the KM estimator, and that the 'distribution to the right' procedure produces the KM estimate.

Supplementary Exercise 4.4 The self-consistent property can also be used with more complicated censoring, such as interval censoring and – as the most extreme case – 'current status' data (e.g., the avalanche dataset) where each observation is either left-censored (dead when extracted) or right-censored (alive when extracted)

EXERCISE: Consider a dataset with 10 observations: the *true* values have no time element, but are (possibly repeated) prime numbers between 1 and 29 inclusive. 6 are left-censored (<10, <16, <18, <21, <26, <28) and 4 are right-censored (>6, >10, >11, and >24.

Analytically, and separately by repeated (iterative) use of the 'self-consistency' principle, arrive at an estimate of $S(t)$.

Hint: You may find the diagram produced by the supplied R code (see website) helpful to visualize the data-intervals.

Start by choosing the support points (here integers) over which the total of probability mass of 1 will be distributed. Try to have these integer values [points of 'support'] be as helpful as possible – include them in (and thus make them contribute to the likelihood of) as many of the data-intervals as possible. In this example, the minimal set of support points has size 3 (note that the 3 points are not unique).

Analytically: write down the likelihood as a function of the magnitudes, θ_1, θ_2, and $\theta_3 = (1 - [\theta_1 + \theta_2])$ of these 'parameters.' Then maximize this with respect to θ_1 and θ_2, say.

Iteratively: Start by strategically selecting 3 probability masses $\{\theta_1^{[0]}, \theta_2^{[0]}, \theta_3^{[0]}\}$ to distribute over the 3 selected support points. This distribution gives you an initial estimate, $S_0(t)$, of the $S(t)$ function. (Out of interest, calculate the Likelihood associated with this $S(t)$).

Then use this $S_0(t)$ as the $S(t)$ in the right hand side of the equation at the beginning of section 4.4.2 to obtain a new estimate, $S_1(t)$ of the $S(t)$ function. (again, out of interest, calculate the Likelihood associated with this new $S(t)$)

Repeat until the estimate of the $S(t)$ function (and the Likelihood) no longer changes.

EXERCISE: Use the supplied R code (or 'roll your own' code) to obtain a NPMLE of the $S(t)$ function in the case of the avalanche data.

[16] Google "Efron distribute to the right Kaplan Meier"

4.4.3 The Nelson-Aalen estimator of $S(t)$

Just as with K-M, divide the entire interval $[0, t]$ into J *narrow* event-containing sub-intervals; ignore the 'non-event-containing' sub-intervals. Sub-interval j is defined by distinct event-time t_j, with n_j at risk just before the event(s) [death(s)] in that interval. (there can be more than 1 event at the same t_j, particularly if time is measured coarsely). The (step-)function $n(t)$ is the number at risk at each time point in $(0, t)$. 'Riskset'$_j$ = the n_j'candidates' for the event(s) at t_j. Suppose s_j survive event-containing sub-interval j, and that the remaining $d_j = n_j - s_j$ do not [the letter d is used here because in many applications, the 'transition' ('event') is from the initial state of 'alive' to the destination state of '$d$ead', but transitions may be desirable or underirable.]

The Nelson-Aalen Estimator uses the same general formula that links the $S(t)$ and $ID(t)$ or $\lambda(t)$ functions:

$$\widehat{S_{NA}(t)} = \exp\Big\{-\int_0^t ID(u)du\Big\} = \exp\Big\{-\int_0^t \lambda(u)du\Big\} = \exp\Big\{-\sum \frac{d_j}{n_j}\Big\}$$

Think of a fitted ID function $ID(t)$ with $\widehat{ID(t)} = 0$ in the non-event-containing sub-intervals of $(0, t)$ and $\widehat{ID(t)} = d/PT = d/(n \times \delta t)$ in each event-containing interval of width δt; thus $\widehat{ID(t)} = d_j/(n_j \times \delta t)$ in event-containing interval j.

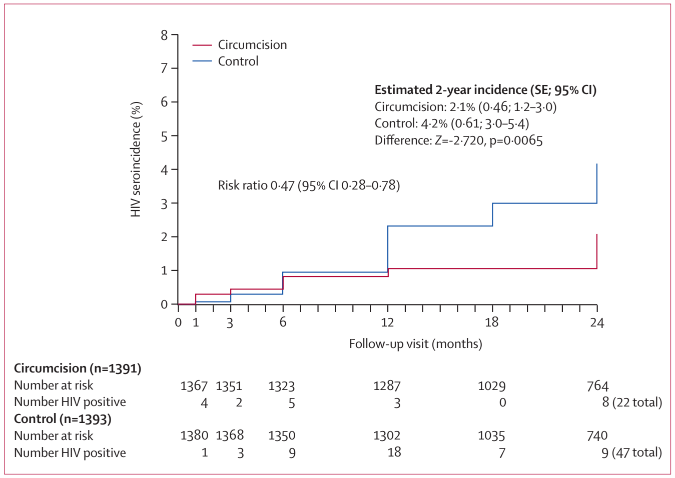
Supplementary Exercise 4.5 (a) Using the $\widehat{ID(t)}$ function just described, evaluate the integral of $\int_0^t \widehat{ID(u)}du$ and use it to obtain the Nelson-Aalen estimator of $S(t)$. (b) Derive the conditions under which the K-M estimator $\prod \frac{s_j}{n_j} = \prod\{1 - \frac{d_j}{n_j}\}$ gives a result that is very close to that of the Nelson-Aalen estimator. (c) Assuming $d_j \sim Poisson(n_j \times \delta t)$, derive an expression for $Var[\widehat{S(t)_{NA}}]$.

4.5 Examples of the Kaplan-Meier method

Example 1 Cf. JUPITER data on the website for course `EPI634`.

The `R` code calls the "canned" routines, but also derives the K-M-based cumulative incidence curves 'from scratch.'

Example 2 Figure 2 below is from the article: "Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial" (Lancet 2007; 369: 643-656). If interested, and if you don't have direct access to the Lancet site, the full article is also available under "resources for rates" in course `EPI634`. There you will also find a companion article for a similar randomized trial, with similar estimates of benefit, carried out in Uganda, and published back to back with the one from Kenya.

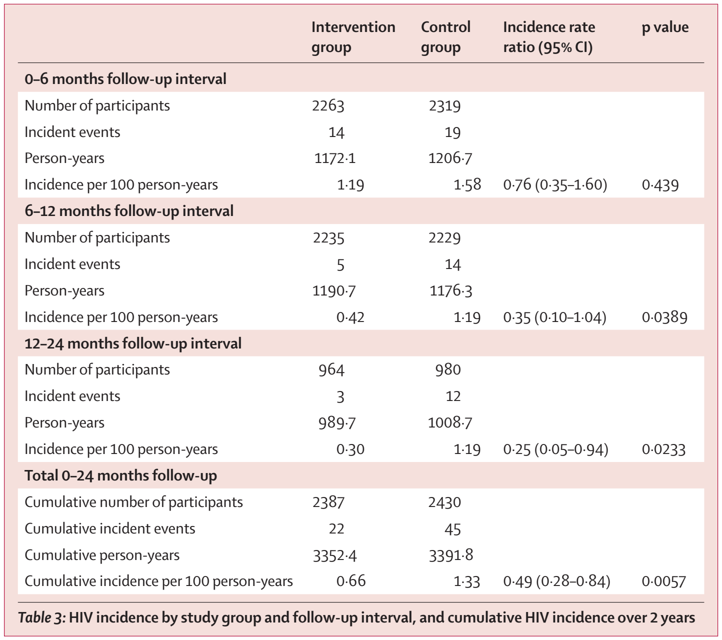


Circumcision (n=1391)						
Number at risk	1367	1351	1323	1287	1029	764
Number HIV positive	4	2	5	3	0	8 (22 total)
Control (n=1393)						
Number at risk	1380	1368	1350	1302	1035	740
Number HIV positive	1	3	9	18	7	9 (47 total)

***Figure 2:* Cumulative HIV seroincidence across follow-up visits by treatment**
Time to HIV-positive status is taken as the first visit when a positive HIV test result is noted. Time is credited as the follow-up visit month. Participants without HIV-positive status are censored at the last regular follow-up visit completed where HIV testing was done, credited specifically as months 1, 3, 6, 12, 18, and 24.

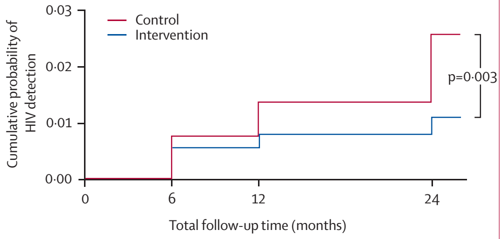
Supplementary Exercise 4.6 Replicate the statistics reported in the insert beginning with the text "Estimated 2-year incidence" in the top right portion of the above Figure 2.

Example 3 The items below are from "Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial," Lancet 2007; 369: 657-666.



	Intervention group	Control group	Incidence rate ratio (95% CI)	p value
0–6 months follow-up interval				
Number of participants	2263	2319		
Incident events	14	19		
Person-years	1172·1	1206·7		
Incidence per 100 person-years	1·19	1·58	0·76 (0·35–1·60)	0·439
6–12 months follow-up interval				
Number of participants	2235	2229		
Incident events	5	14		
Person-years	1190·7	1176·3		
Incidence per 100 person-years	0·42	1·19	0·35 (0·10–1·04)	0·0389
12–24 months follow-up interval				
Number of participants	964	980		
Incident events	3	12		
Person-years	989·7	1008·7		
Incidence per 100 person-years	0·30	1·19	0·25 (0·05–0·94)	0·0233
Total 0–24 months follow-up				
Cumulative number of participants	2387	2430		
Cumulative incident events	22	45		
Cumulative person-years	3352·4	3391·8		
Cumulative incidence per 100 person-years	0·66	1·33	0·49 (0·28–0·84)	0·0057

Table 3: **HIV incidence by study group and follow-up interval, and cumulative HIV incidence over 2 years**



Cases of HIV/total participants

	0	6	12	24
Intervention	0/2474	14/2387	5/2274	3/964
Control	0/2522	19/2430	14/2279	12/980

Figure 2: **Kaplan-Meier cumulative probabilities of HIV detection by study group**

Actual visits grouped by the three scheduled visits at 6 months, 12 months, and 24 months after enrolment. The cumulative probabilities of HIV infection were 1·1% in the intervention group and 2·6% in the control group over 24 months.

Supplementary Exercise 4.7 Comment on the appropriateness of (i) the term "*Cumulative incidence per 100 person-years*" in the last row of Table 3 (ii) using a single incidence (hazard) rate ratio of 0.49 for the full 2 years, and in the abstract, reporting that the estimated efficacy of intervention was 51%.

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*[17], 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.[18] 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.

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. 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. 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 $(t, t + \delta t)$,

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

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/$Person$ is 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

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

[17] 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.

[18] 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."

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), 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 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").[19] If the mean number 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

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

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

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 1year 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.[20] 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 2year 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

$$L = \exp[-(2300 \times 0.5 + 2215 \times 0.5 + 968 \times 1.0)\lambda]$$
$$\times$$
$$(1 - \exp[-0.5\lambda])^{19} \ \times \ (1 - \exp[-0.5\lambda])^{14} \ \times \ (1 - \exp[-1.0\lambda])^{12}$$

Supplementary Exercise 5.1. (i) Maximize L with respect to λ. (ii) 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?

[20] 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.

5.4 Cum. survival probability as fn. of rate parameter

We saw this in BIOS601 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 $0, 1, 2, \dots$ 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 in first term 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 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 a) 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 5.2. 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.

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 ($\underline{d}$eaths), 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 5.3. 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.

6 Time

6.1 When do we start the clock?

Examples JH has dealt with include the analysis of longevity of

- The Titanic survivors, where the two time scales are (i) age (years elapsed since birth) and (ii) 'survivor-time', the years elapsed since the April 15, 1912 sinking;
- Oscar nominees, where the two time scales are (i) age and (ii) nominee-time', the years elapsed since first being nominated for an Oscar;
- Nobel Prize nominees, where the two time scales are (i) age and (ii) 'nominee-time', the years elapsed since first being nominated for a Nobel Prize;
- Jazz musicians, where the two time scales are (i) age and (ii) performer-time', the years elapsed since first becoming a jazz musician;
- Popes versus artists;
- Baseball Hall of Famers versus players who were nominated by not inducted;
- Rock Stars who become famous early versus later (or not at all).

  For more details on these examples, see `bios601/Epidemiology2/`

For more on the choice of time scale, Google "Multiple time scales in survival analysis." or find the articles that cite the 1979 Applied Statistics article by Farewell and Cox "A note on multiple time scales in life testing."

There is also the interesting article *The two-way proportional hazards mode*l by Efron in J. R. Statist. Soc. B (2002) 64, Part 4, pp. 899-909, applied to "patient histories in a study of heart transplant recipients treated at the Stanford Medical Center between 1980 and 1996; some 110 of the patients suffered a *serious bacterial infection*, their infection times ranging from a few *days after transplantation* to nearly 9 years, these being the observed lifetimes that would usually be featured in a proportional hazards analysis of the infection process. In this case, however, the investigators' *main interest centred on calendar date*: was the *incidence rate* of bacterial infections *declining over the course of the study?* Incidence is itself a hazard rate, in the simplest situation the number of new cases per eligible subject per unit time, and it is natural to answer the question with a hazard rate analysis."

6.2 Age-specific rates

"*To ignore this variation [of incidence and mortality rates with age] runs the risk that comparisons between groups will be seriously distorted, or confounded, by differences in age structure.*"

It's good to have a few handy real examples of *age-confounding* that are easily understood by non-statisticians. Two immediately come to mind (i) the overall death rate is higher in Canada than Ethiopia (ii) the higher death rate among non-smokers in a 20-year follow-up study of smokers and non-smokers [Does Smoking Improve Survival? `www.whfreeman.com/statistics/ips/eesee4/eesees4.htm`; this is also described in chapter 1 of Rothman 2002, with finer age-categories]

"*For longer studies it will be necessary to take account of changing age during the study, and to treat age properly - as a time scale. This scale is then divided into bands and a separate estimate of the rate is made within each age band as described in Chapter 5. In this latter analysis, a subject can pass through several age bands during the course of the study.*"

Not only can a subject pass through several age bands but she can also change from one 'exposure' category to another – as in the Oscars exercise.

6.3 The expected number of failures

"*One reason for subdividing the total follow-up experience of a cohort into age bands is to determine whether the observed number of failures is more or less than we might have expected. Since mortality and incidence rates usually increase quite sharply with age, the distribution of person years observation between age bands is an extremely important determinant of the number of events we would expect to observe.*"

It is not clear what is the basis for the "expectation" i.e., whether it is a 'what if' comparison against external rates, or an internal one against the rates in a comparison group constructed and followed by the investigators. One can think of the 'expected number' of 16.77 cases in exercise 6.3 as the number one would expect in a scaled-down version of England and Wales (E&W), scaled down to the same sample size (974 women) followed for the same cell-specific numbers of person years as those shown in Table 6.4. In other words, it as as thought one had

974 treated by HRT	974 from E&W, same age & follow-up, untreated
15 cases	16.7 cases

Of course, the fact that the 16.7 is based on observed rates in the whole of

E&W means that it is not subject to the same degree of random variation as is the number of cases in the actual cohort. With this solid a basis for it, the expected number is usually taken to be a constant, so only one standard error (SE) is involved in the 15 vs. 16.7 comparison – the one associated with the 15.

"The expected number of cases, as calculated above, is not quite the same as the expected number in the usual statistical sense. The latter cannot depend upon the outcome of the study, but the former does."

C&H are saying that the numbers of Woman-years in the second column of Table 6.4 are random variables: they would not have been known ahead of time. For some 15 women – the 15 being a random variable – the follow-up was terminated by the event of interest. Likewise, any terminations for other reasons might also be unpredictable ahead of time. However, if these are not related to the person's probability of a future event, they don't have a great influence on the sampling behaviour of the estimators of interest.

6.4 Lexis diagrams

`en.wikipedia.org/wiki/` Wilhelm Lexis (1837-1914) was an eminent German statistician, economist, and social scientist and a founder of the interdisciplinary study of insurance.

The "Lexis diagram", in which lifelines are displayed as 45-degree lines on a grid with age on the vertical axis and calendar year on the horizontal axis, is very helpful in epidemiology, and in survival analysis with 2 time scales.

The `Epi` package for `R` has several functions that make it easy to convert the data of the type shown in Table 6.2 into the person-year segments shown Figure 6.3. Previously, this was a very laborious computing process.

Once we have the tabulated person years and cases in each Lexis rectangle (the cells don't have to be square), we can calculate the expected number of cases if a specified set of external rates applied, or make internal rectangle-by-rectangle comparisons, and thus a summary of these comparisons. We can also use them to fit (Poisson) regression models for rates.

Here is the `R` code, and some of its output, for the data in C&H Table 6.2.

```
library(Epi)

id = c(1,2,3,4);
yr.birth = c(1904,1924,1914,1920);
yr.entry = c(1943,1948,1945,1948);
yr.exit  = c(1952,1955,1961,1956);
fail = c(0, 1, 0, 0) );

ds=data.frame(id, yr.birth, yr.entry, yr.exit, fail); ds

  id yr.birth yr.entry yr.exit fail
1  1     1904     1943    1952    0
2  2     1924     1948    1955    1
3  3     1914     1945    1961    0
4  4     1920     1948    1956    0


# Define as Lexis object with timescales calendar time and age

Lexis <- Lexis( entry = list( calendar.year = yr.entry ),
                exit  = list( calendar.year = yr.exit, age = yr.exit - yr.birth ),
         exit.status  = fail,
                data  = ds )
Lexis

  calendar.year age lex.dur lex.Cst lex.Xst lex.id id yr.birth yr.entry yr.exit fail

1          1943  39       9       0       0      1  1     1904     1943    1952    0
2          1948  24       7       0       1      2  2     1924     1948    1955    1
3          1945  31      16       0       0      3  3     1914     1945    1961    0
4          1948  28       8       0       0      4  4     1920     1948    1956    0

# Default plot of follow-up

plot(Lexis)

# With a grid and deaths as endpoints

plot(Lexis, grid=0:5*5, col="black" )
points(Lexis, pch=c(NA,16)[Lexis$lex.Xst+1] )

# With a lot of bells and whistles:  [ *** SEE PLOT NEXT PAGE *** ]

plot(Lexis, grid=0:20*5, col="black", xaxs="i", yaxs="i",
      xlim=c(1940,1965), ylim=c(20,50), lwd=3, las=1 )
points(Lexis, pch=c(NA,16)[Lexis$lex.Xst+1], col="red", cex=1.5 )

# Split time along two time-axes

L2 = splitLexis(Lexis,breaks=seq(1940,1965,5),
   time.scale="calendar.year")
L2 = splitLexis(L2,   breaks=seq(20,50,5), time.scale="age" )
str( L2 )
```

```
L2

   lex.id calendar.year age lex.dur lex.Cst lex.Xst id yr.birth yr.entry yr.exit fail
1       1          1943  39       1       0       0  1     1904     1943    1952    0
2       1          1944  40       1       0       0  1     1904     1943    1952    0
3       1          1945  41       4       0       0  1     1904     1943    1952    0
4       1          1949  45       1       0       0  1     1904     1943    1952    0
5       1          1950  46       2       0       0  1     1904     1943    1952    0
6       2          1948  24       1       0       0  2     1924     1948    1955    1
7       2          1949  25       1       0       0  2     1924     1948    1955    1
8       2          1950  26       4       0       0  2     1924     1948    1955    1
9       2          1954  30       1       0       1  2     1924     1948    1955    1
10      3          1945  31       4       0       0  3     1914     1945    1961    0
11      3          1949  35       1       0       0  3     1914     1945    1961    0
12      3          1950  36       4       0       0  3     1914     1945    1961    0
13      3          1954  40       1       0       0  3     1914     1945    1961    0
14      3          1955  41       4       0       0  3     1914     1945    1961    0
15      3          1959  45       1       0       0  3     1914     1945    1961    0
16      3          1960  46       1       0       0  3     1914     1945    1961    0
17      4          1948  28       2       0       0  4     1920     1948    1956    0
18      4          1950  30       5       0       0  4     1920     1948    1956    0
19      4          1955  35       1       0       0  4     1920     1948    1956    0
```

```
# Tabulate the cases and the person-years

summary( L2 )

tapply( status(L2,"exit")==1, list( timeBand(L2,"age","left"),
          timeBand(L2,"calendar.year","left") ), sum )

   1940 1945 1950 1955 1960
20   NA    0   NA   NA   NA
25   NA    0    0   NA   NA
30   NA    0    1   NA   NA
35    0    0    0    0   NA
40    0    0    0    0   NA
45   NA    0    0    0    0

tapply( dur(L2),  list( timeBand(L2,"age","left"),
          timeBand(L2,"calendar.year","left") ), sum )

   1940 1945 1950 1955 1960
20   NA    1   NA   NA   NA
25   NA    3    4   NA   NA
30   NA    4    6   NA   NA
35    1    1    4    1   NA
40    1    4    1    4   NA
45   NA    1    2    1    1
```

```
> summary( L2 )

Transitions:
     To
From   0 1 Records:  Events:  Risk time:
   0 18 1       19         1          40

Rates:
     To
From 0    1 Total
   0 0 0.02  0.02
```

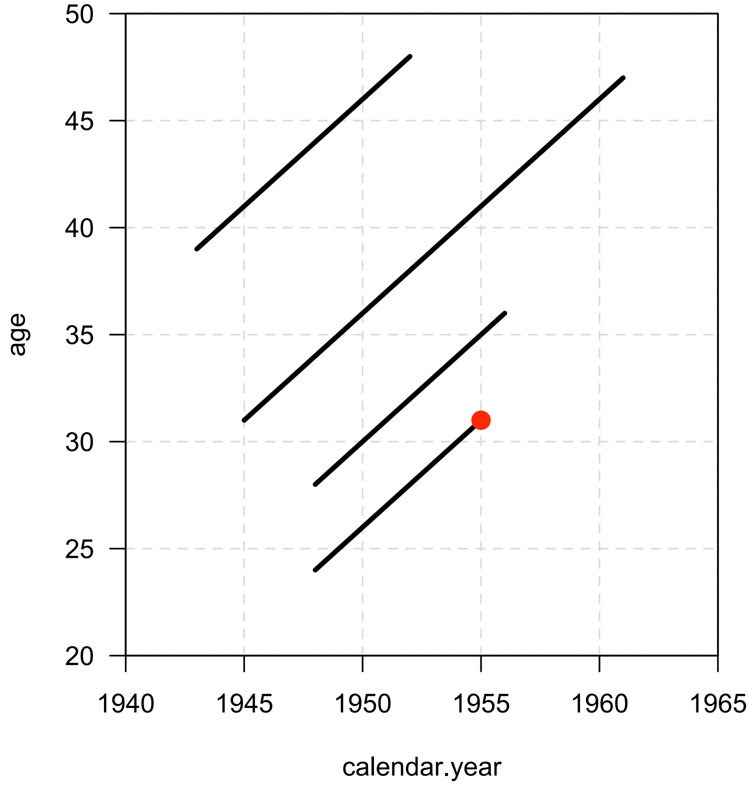


Figure 4: Lexis Diagram, from `Epi` package in `R`

Supplementary Exercise 6.1. Death rates in those who survived the sinking of the Titanic vs. in the sex-and age-matched US general population, together with some other investigations

Under 'For Person-Years Analyses' in Resources for 'Fitting Models to Grouped Data [B & D vol II, ch4]' in the BIOS602 website you will find (a) the Titanic longevity data set (b) USA death rates (within 5 x 5 rectangles, called '*quinquinquennia*') from the Berkeley Mortality Database.[21] You will also find some R code that uses the Epi package to create – for each passenger – the durations in and exit status from each quinquinquennium, then aggregates these over all the persons traversing each quinquinquennium, etc.

1. Convert each survivor's record into the experience in the (age, period) quinquinquennia traversed, i.e the number of years spent in the rectangle, and the status (e.g., $d = 0$ if alive, 1 if dead) at the end of these years. Rather than program the calculations from scratch, two possibilities are `http://epi.klinikum.uni-muenster.de/pamcomp/pamcomp.html` – which some people used last year – and the `R` 'Epi' package `http://staff.pubhealth.ku.dk/∼bxc/Epi/` The key functions in the latter are `Lexis` (and associated plotting functions) and `splitLexis`, which, when applied twice, calculates the time spent, and exit status from each quinquinquennium. The 'bogus example' in the documentation of the `splitLexis` function illustrates these, while the example on the notes for C&H chapter 6 shows the application to the 4-person cohort used in that chapter.

2. How much higher/lower is the *set* of age-specific death rates for male Titanic survivors than that for the general US population? for female survivors? Answer in two ways: first, calculate sex-specific observed/expected ratios, where the numerator is the total number of deaths observed in the sex-specific cohort, and the denominator is the sum of the expected numbers of deaths in these cells, using the USA age-sex-period death rates; second, calculate sex-specific Mantel-Haenszel summary incidence ratios (Rothman terminology) or incidence density ratios (Miettinen terminology) or mortality rate ratios (everyone's terminology), using age and period as 'strata.'[22] Assume that each of the USA death rates is based on a denominator of one million person years.[23] Assume that the death rates after 1995 are the same as those in 1990-95.

3. 'On average,' [24], for the age-span 40-90 in the period 1990-1995, how much higher are the USA age-specific male death rates in males than females? Answer by plotting the log of the male:female death rate ratio vs age, (or the two separate sets of log-death-rates on the same graph), and taking some 'typical' value for the ratio. Are you comfortable giving a single ratio? i.e., is the mortality-rate-ratio (M:F) reasonably constant over that age-span?

4. The previous question refers to cross-sectional rates, i.e., those in a specified *period*.[25] On average, over the age-span 40-90 in the 1900 *birth-cohort*, how much higher are the USA age-specific death rates in males than females? Answer by plotting the log of the male:female death rate ratio vs age, (or the two separate sets of log-death-rates on the same graph), and taking some 'typical' value for the ratio. Are you comfortable giving a single ratio? i.e., is the mortality-rate-ratio (M:F) *reasonably* constant over that age-span?

5. For the age-span 40-90, in a single number describe how much age-and specific death rates have fallen over the 20th century (the changes may be more subtle that this, so your answer will necessarily be a simplification).

6. For the Titanic survivors, was there a gradient in mortality rates across the 3 passenger classes?

Supplementary Exercise 6.2. Mortality of performers while in the 'still hoping to win' vs in the 'already a winner' state

1. Divide the performer-years into those spent as Oscar nominees and as Oscar winners and then subdivide these into quinquinquennia.

2. Compare the death rates in the performer-years spent as nominees versus those spent as winners. Do so using both 'adjusted' expected numbers and purely-internal comparisons.

[21] .] This site, `http://www.demog.berkeley.edu/∼bmd/index.html`, contains historical lifetable and death rate data for the USA and other countries.

[22] As is illustrated in equation 8-5 in Rothman 2002, the formula is

$$\frac{\sum_{strata}(no.\ of\ cases,\ index\ category) \times (py,\ ref.\ category)/(py\ in\ stratum)}{\sum_{strata}(no.\ of\ cases,\ ref.\ category) \times (py,\ index\ category)/(py\ in\ stratum)}$$

[23] If the ratio of the amount of experience in the ref. category to that in the index category goes to infinity, the M-H summary ratio converges to $\sum_{strata} O / \sum_{strata} E = O/E$.

[24] Even if the average is not representative.

[25] Cross-sectional rates are what are used to make 'current' or 'period' lifetables, by far the more common type of lifetable.