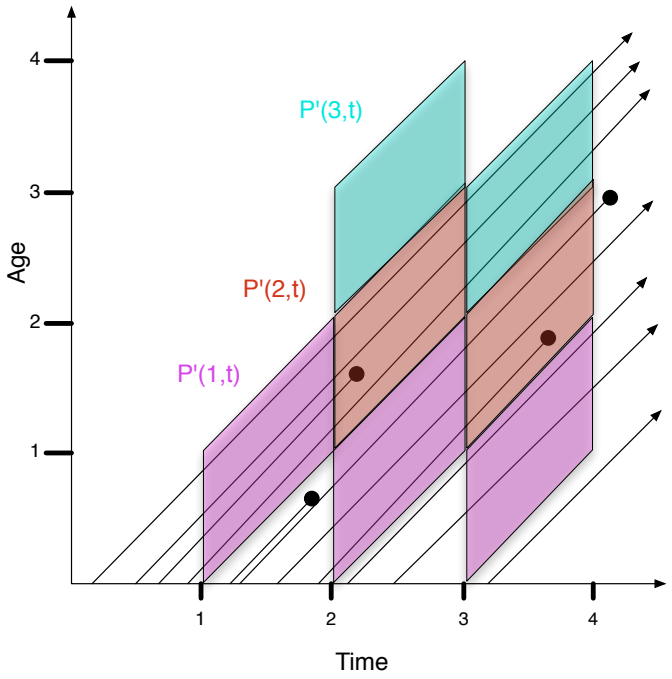
BS3B
Statistical Lifetime-Models

David Steinsaltz[1]
University of Oxford
Based on early editions by Matthias Winkel and Mary Lunn



HT 2010 (updated HT 2014)

[1]University lecturer at the Department of Statistics, University of Oxford

BS3b
Statistical Lifetime-Models

David Steinsaltz – 16 lectures HT 2010
steinsal@stats.ox.ac.uk

Prerequisites

Part A Probability, Part A Statistics and Part B Applied Probability are prerequisites.

Website: http://www.steinsaltz.me.uk/BS3b/BS3b.html

Aims

Statistical Lifetime-Models follows on from Applied Probability. Models introduced there are examined in the first part of the course more specifically in a life insurance context where transitions typically model the passage from 'alive' to 'dead', possibly with intermediate stages like 'loss of a limb' or 'critically ill'. The aim is to develop statistical methods to estimate transition rates and more specifically to construct life tables that form the basis in the calculation of life insurance premiums.

We will then move on to survival analysis, which is widely used in medical research, in addition to insurance, in which we consider the effect of covariates and of partially observed data. We also explain demographic concepts, and how life tables are adapted to the context of changing mortality rates.

Synopsis

Survival models: general lifetime distributions, force of mortality (hazard rate), survival function, specific mortality laws, the single decrement model, curtate lifetimes, life tables, period and cohort.

Estimation procedures for lifetime distributions: empirical lifetime distributions, censoring, Kaplan-Meier estimate, Nelson-Aalen estimate. Parametric models, accelerated life models including Weibull, log-normal, log-logistic. Plot-based methods for model selection. Proportional hazards, partial likelihood, semiparametric estimation of survival functions, use and overuse of proportional hazards in insurance calculations and epidemiology.

Two-state and multiple-state Markov models, with simplifying assumptions. Estimation of Markovian transition rates: Maximum likelihood estimators, time-varying transition rates, census approximation. Applications to reliability, medical statistics, ecology.

Graduation, including fitting Gompertz-Makeham model, comparison with standard life table: tests including chi-square test and grouping of signs test, serial correlations test; smoothness.

Exercises and Classes

Classes will be held on Fridays. There will be four sessions: 10-11, 11-12, 2-3, 3-4. Class assignments will be available on Minerva at https://minerva.stats.ox.ac.uk.

Scripts are to be handed in by Mondays 4pm in the Department of Statistics.

The scope of exercises goes significantly beyond that of exam questions in many cases, but understanding the exercises is essential to coping with the variety of exam questions that might come up. There is a great range of difficulty in the exercises, and most students should find at least some of the exercises very challenging. Try all of them, but don't spend hours and hours on questions if you are not making any progress.

Try to start solving exercises when you get the problem sheet, not the day before you have to hand in your solutions. This allows you to have second attempts at exercises that you can't solve straight away.

Lecture notes are meant to be useful when solving exercises. You may use any result from the lectures, except where the contrary is explicitly stated.

Reading

There are lots of good books on survival analysis. Look for one that suits you. Some pointers will be given in the lecture notes to readings that are connected, but look in the index to find topics that confuse you and/or interest you.

The actuarial material in the course is modeled on the CT4 Core Reading from the Institute of Actuaries.

CT4 Models Core Reading. Faculty & Institute of Actuaries

This is the core reading for the actuarial professional examination on survival models. In some places, the approach is more practically oriented and often placed in an insurance context, whereas the course is more academic and not only oriented towards insurance applications. All in all, this is the main reference for about half the course. It is available for about £21.50 from the Institute of Actuaries on Worcester Street. (A few college libraries have it.)

D.R. Cox and D. Oakes: *Analysis of Survival Data.* Chapman & Hall (1984)

This is *the* classical text on survival analysis. The presentation is concise, but gives a broad view of the subject. The text contains exercises. This is the main reference for about half the course. It contains also much more related material beyond the scope of the course.

H.U. Gerber: *Life Insurance Mathematics.* 3rd edition, Springer (1997)

The presentation is concise. Only three chapters are relevant. Chapter 2 gives an introduction to lifetime distributions, Chapter 7 discusses the multiple decrement model and Chapter 11 estimation procedures for lifetime distributions. The remainder combines the ideas with the interest rate theory of BS4.

Klein & Moeschberger: *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd edition, Springer (2003)

This is an excellent source for a lot of the survival analysis topics, particularly censoring and truncation, and the Kaplan-Meier and Nelson-Aalen estimators. Lots of terrific examples.

Contents

Lecture 1

Introduction: Survival Models

1.1 Early life tables

In one of the earliest treatises on probability George Leclerc Buffon considered the problem of finding the fundamental unit of risk, the smallest discernible probability. He wrote that "all fear or hope, whose probability equals that which produces the fear of death, in the moral realm may be taken as unity against which all other fears are to be measured." [Buf77, p. 56] In other words, because no healthy man in the prime of life (he argued) attends to the risk that he may die in the next twenty-four hours, Buffon considered that events with this probability could be treated as negligible; after all, "since the intensity of the fear of death is a good deal greater than the intensity of any other fear or hope," any other risk of equivalent probability of a less troubling event — such as winning a lottery — would leave a person equally indifferent. He decided that the appropriate age to consider for a man to be in the prime of health was 56 years. But what is that probability, that a 56 year old man dies in the next day?

To answer this, Buffon turned to mortality tables. A colleague (one M. Dupré of Saint-Maur) assembled the registers of 12 rural parishes and 3 parishes of Paris, in which 23,994 deaths were recorded. The ages at death were all recorded, so that he knew that 174 of the deaths were at age 56; that is, between the 56th and 57th birthdays.[1] Our naïve estimator for the probability of an event is

$$\text{probability of occurrence} = \frac{\text{number of occurrences}}{\text{number of opportunities}}.$$

The number of occurrences of the event (death of an individual aged 56) is observed to be 174. But what about the denominator? The number of "opportunities" for this event is just the number of individuals in the population at the appropriate age. The most direct way to determine this number would be a time-consuming census. Buffon's approach (and that of other 17th and 18th creators of such life tables) depended upon the following implicit logic: Suppose the population is stable, so that the same number of people in each age group die each year. Since every person dies at some time (it is believed), the total number of people in the population who live to their 56th birthday will be exactly the same as the number of people observed to have died *after* their 56th birthday in the particular year under observation, which happens to be 5031. The probability of dying in one day may then be estimated as

$$\frac{1}{365} \times \frac{174}{5031} \approx \frac{1}{10000},$$

[1] Actually, Buffon's statistical procedure was a bit more complicated than this. The recorded numbers of deaths at ages 55,56,57,58,59,60 were 280,130,129,182,90,534 respectively. Buffon observed that the priests ("particularly the country priests") were likely to record round numbers for the age at death, rather than the exact age — which they may not know anyway. He thus decided that it would make more sense to smooth (as statisticians would call the procedure today) or **graduate** (as actuaries call it) the data. We will learn about graduation in Lecture.

and Buffon proceeds to reason with this estimate.

From this elementary exercise we see that:

- Mortality probabilities can be estimated as the ratio of the number of deaths to the number of individuals "at risk".
- The numerator (the number of deaths) is usually straightforward to determine.
- The denominator (the number at risk) can be challenging.
- Mortality can serve as a model for thinking about risks (and opportunities) more generally, for events happening at random times.
- You don't get very far in thinking about mortality and other risks without some sort of theoretical model.

The last claim may require a bit more elucidation. What would a naïve, empirical approach to life tables look like? Given a census of the population by age, and a list of the ages at death in the following year, we could compute the proportion of individuals aged x who died in the following year. This is merely a free-floating fact, which could be compared with other facts, such as the measured proportion of individuals aged x who died in a different year (or at a different age, or a different place, etc.) If you want to talk about a probability of dying in that year (for which the proportion would serve as an estimate), this is a theoretical construct, which can be modelled (as we will see) in different ways. Once you have a probability model, this allows you to pose (and perhaps answer) questions about the probability of dying in a given day, make predictions about past and future trends, and isolate the effect of certain medications or life-style changes on mortality.

There are many different kinds of problems for which the same survival analysis statistics may be applied. Some examples which we will consider at various points in this course are:

- Time to failure of a machine with multiple internal components.
- Time from infection until a subject shows signs of illness.
- Time from starting to try to conceive a baby until a woman is pregnant.
- Time until a person diagnosed with (and perhaps treated for) a disease has a recurrence.
- Time until an unmarried couple marries or separates.

Often, though, we will use the term "lifetime" to represent any waiting time, along with its attendant vocabulary: survival probability, mortality rate, cause of death, etc.

1.2 Basic statistical methods for lifetime distributions

In Table 1.1 we see the estimated ages at death for 103 tyrannosaurs, from four different species, as reported in [ECIW06]. Let us treat them here as a single population.

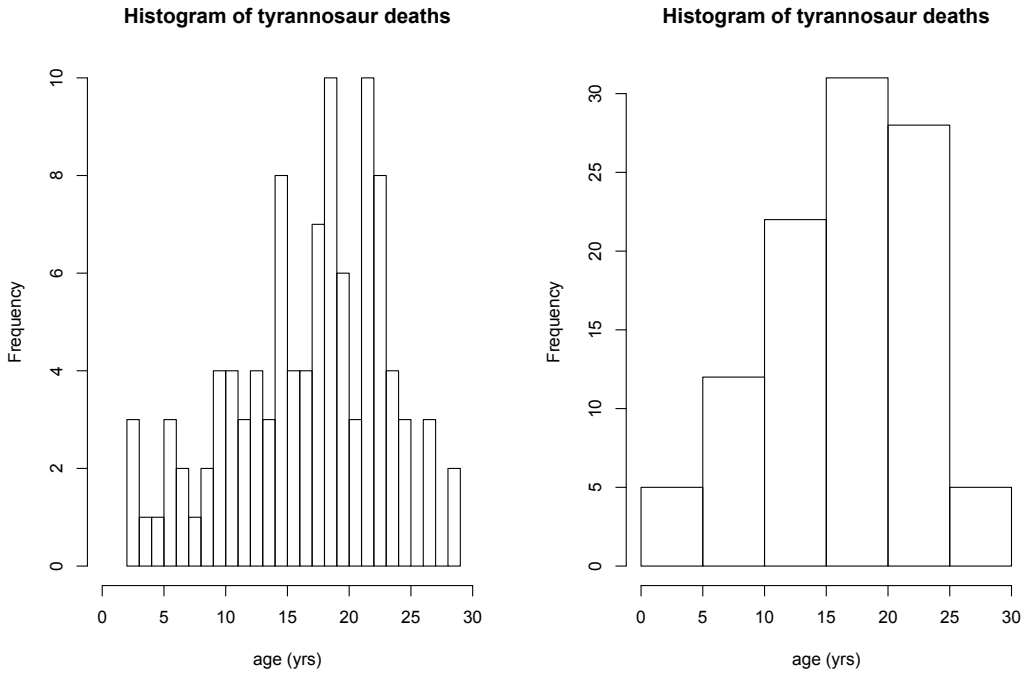
A. sarcophagus	2,4,6,8,9,11,12,13,14,14,15,15,16,17,17,18,19,19,20,21,23,28
T. rex	2,6,8,9,11,14,15,16,17,18,18,18,18,18,19,21,21,21, 22,22,22,22,22,22,23,23,24,24,28
G. libratus	2,5,5,5,7,9,10,10,10,11,12,12,12,13,13,14,14,14,14,14, 15,16,16,17,17,17,18,18,18,19,19,19,20,20,21,21,21,21,22
Daspletosaurus	3,9,10,17,18,21,21,22,23,24,26,26,26

Table 1.1: 103 estimated ages of death (in years) for four different tyrannosaur species.

In Part A Statistics you learned to do the following:

1.2.1 Plot the data

The most basic thing you can do with any data is to sort the observations into bins of some width Δ, and plot the histogram, as in Figure 1.1). This does not presuppose any model.



(a) Narrow bins

(b) Wide bins

Figure 1.1: Histogram of tyrannosaur mortality data from Table 1.1.

1.2.2 Fit a model

Suppose we believe the list of lifetimes to be i.i.d. samples from a fixed (unknown) distribution. We can then use the data to determine which distribution it was that generated the samples.

In Part A statistics you learned parametric maximum likelihood estimation. Suppose the unknown distribution is believed to be one of a family of distributions that is indexed by a possibly multivariate (k-dimensional) parameter $\lambda \in \Lambda \subset \mathbb{R}^k$. That is — taking just the case of data from a continuous distribution — the distribution of the independent observations has density $f(T;\lambda)$ at the point T, if the true value of the parameter is λ. Suppose we have observed n independent lifetimes $T_1, \ldots, T_n$. We define the log-likelihood function to be the (natural) log of the density of the observations, considered as a function of the parameter. By the assumption of independence, this is

$$\ell_{T_1,\ldots,T_n}(\lambda) = \ell_{\mathbf{T}} := \sum_{i=1}^{n} \ln f(T_i;\lambda). \tag{1.1}$$

(We use $\mathbf{T}$ to represent the vector $(T_1, \ldots, T_n)$.) The maximum likelihood estimator (MLE) is simply the value of λ that makes this as large as possible:

$$\hat{\lambda} = \hat{\lambda}(\mathbf{T}) = \hat{\lambda}(T_1, \ldots, T_n) := \arg\max_{\lambda\in\Lambda} \prod_{i=1}^{n} f(T_i;\lambda). \tag{1.2}$$

Notice the nomenclature: $\max_{\lambda\in\Lambda} f(\lambda)$ picks the maximal value in the range of f, $\arg\max_{\lambda\in\Lambda} f(\lambda)$ picks the λ-value in the domain of f for which this maximum is attained.

The most basic model for lifetimes is the exponential. This is the "memoryless" waiting-time distribution, meaning that the remaining waiting time always has the same distribution, conditioned on the event not having occurred up to any time t. This distribution has a single parameter ($k = 1$) μ, and density

$$f(\mu;T) = \mu e^{-\mu T}.$$

The parameter μ is chosen from the domain $\Lambda = (0,\infty)$. If we observe independent lifetimes $T_1, \ldots, T_n$ from the exponential distribution with parameter μ, and let $\bar{T} := n^{-1}\sum_{i=1}^{n} T_i$ be the average, the log likelihood is

$$\ell_{\mathbf{T}}(\mu) = \sum_{i=1}^{n} \ln\left(\mu e^{-\mu T_i}\right) = n\left(\ln\mu - \bar{T}\mu\right),$$

which has maximum at $\hat{\mu} = 1/\bar{T} = n/\sum T_i$. This is an example of what we will see to be a general principle:

$$\text{Estimated rate} = \frac{\#\text{ events}}{\text{total time at risk}}. \tag{1.3}$$

In some cases we will be thinking of the time as random, in other cases the number of events, but the formula (1.3) remains. The challenge will be to estimate the number of events and the total time in a way that they correspond to the same time period and the same population, since they are often estimated from different data sources and timed in different ways.

For large n, the estimator $\hat{\lambda}(T_1, \ldots, T_n)$ is approximately normally distributed, under some regularity conditions, and it has some other optimality properties (finite-sample and asymptotic). This allows us to construct approximate confidence intervals/regions to indicate the precision of maximum likelihood estimates. Specifically, for

$$\hat{\lambda} \sim \mathcal{N}\left(\lambda, (I(\lambda))^{-1}\right), \quad \text{where } I_{j_1 j_2}(\lambda) = -\mathbb{E}\left[\frac{\partial^2}{\partial\lambda_{j_1}\partial\lambda_{j_2}} \sum_{i=1}^{n} \ln(f(T_i;\lambda))\right] = -\mathbb{E}\left[\frac{\partial^2 \ell_{\mathbf{T}}(\lambda)}{\partial\lambda_{j_1}\partial\lambda_{j_2}}\right]$$

are the entries of the Fisher Information matrix. Of course, we generally don't know what λ is — otherwise, we probably would not be bothering to estimate it! — so we may approximate

the Information matrix by computing $I_{j_1 j_2}(\hat{\lambda})$ instead. Furthermore, we may not be able to compute the expectation in any straightforward way; in that case, we use the principle of Monte Carlo estimation: We approximate the expectation of a random variable by the average of a sample of observations. We already have the sample $T_1, \ldots, T_n$ from the correct distribution, so we define the *observed information matrix*

$$J_{j_1 j_2}(\lambda, T_1, \ldots, T_n) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln f(T_i; \lambda)}{\partial \lambda_{j_1} \partial \lambda_{j_2}}.$$

Again, we may substitute $J_{j_1 j_2}(\hat{\lambda}, T_1, \ldots, T_n)$, since the true value of λ is unknown. Thus, in the case of a one-dimensional parameter (where the covariance matrix is just the variance and the matrix inverse $(I(\hat{\lambda}))^{-1}$ is just the multiplicative inverse in $\mathbb{R}$), we obtain

$$\left[\hat{\lambda} - 1.96\sqrt{\frac{1}{I(\hat{\lambda})}}, \hat{\lambda} + 1.96\sqrt{\frac{1}{I(\hat{\lambda})}}\right]$$

as an approximate 95% confidence interval for the unknown parameter λ.

In the case of the exponential model, we have

$$\ell''_{\mathbf{T}}(\mu) = -\frac{n}{\mu^2},$$

so that the standard error for $\hat{\mu}$ is $\mu/\sqrt{n}$, which we estimate by $\hat{\mu}/\sqrt{n}$. For the tyrannosaur data of Table 1.1, we have

$$\begin{aligned} \bar{T} &= 16.03, \\ \hat{\mu} &= 0.062, \\ SE_{\hat{\mu}} &= 0.0061, \\ 95\% \text{ confidence interval for } \hat{\mu} &= (0.050, 0.074). \end{aligned}$$

Aside: In the special case of exponential lifetimes, we can construct exact confidence intervals, since we know the distribution of $n/\hat{\mu} \sim \Gamma(n, \mu)$, so that $2n\mu/\hat{\mu} \sim \chi^2_{2n}$ allows to use χ^2-tables.

Is the fit any good? We have various standard methods of testing goodness of fit — we discuss an example in section 1.2.3 — but it's pretty easy to see by eye that the histograms in Figure 1.1 aren't going to fit an exponential distribution, which is a declining density, very well. In Figure 1.2 we show the empirical (observed) cumulative distribution of tyrannosaur deaths, together with the cdf of the best exponential fit, which is obviously not a very good fit at all.

We also show (in green) the fit to a class of distribution which is an example of a larger class that we will meet later, called the "Weibull" distributions. Instead of the exponential cdf $F(t) = 1 - e^{-\mu t}$, suppose we take $F(t) = 1 - e^{-\alpha t^2}$. Note that if we define $Y_i = T_i^2$, we have

$$P(Y_i \leq y) = P(T_i \leq \sqrt{y}) = 1 - e^{-\alpha y},$$

so Y_i is actually exponentially distributed with parameter α. Thus, the MLE for α is

$$\hat{\alpha} = \frac{n}{\sum T_i^2}.$$

We see in Figure 1.2 that this fits much better than the exponential distribution.

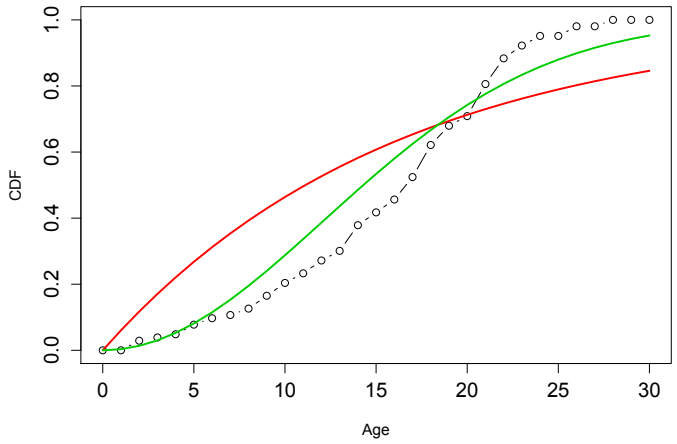


Figure 1.2: Empirical cumulative distribution of tyrannosaur deaths (circles), together with cdf of exponential fit (red) and Weibull fit (green).

1.2.3 Significance test

The maximum-likelihood approach is optimal in many respects for picking the correct parameter based on the observed data, **under the assumption that the observed data did actually come from a distribution in the appointed parametric family**. But did they? We already looked at a plot, Figure 1.2, comparing the fit cdf to the observed cdf. The Weibull fit was clearly better. But how much better?

One way to answer this question is to apply a significance test. We start with a set of distributions H_1, such that we know that it includes the true distribution (for instance, the set of all distributions on $(0, \infty)$, and a null hypothesis $H_0 \subset H_1$, and we wish to test how plausible the observations are as a sample from H_0, rather than from the alternative hypothesis $H_1 \setminus H_0$. The standard parametric procedure is to use a χ^2 goodness of fit test, based on the statistic

$$X^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \sim \chi^2_{m-k-1} \qquad \text{approximately, under } H_0, \tag{1.4}$$

where m is the number of bins (e.g. from your histogram), but merged to satisfy size restrictions, and k the number of parameters estimated. O_j is the random variable modelling the number observed in bin j, E_j the number expected under maximum likelihood parameters. To justify the approximate distribution for the test statistic, we require that at most 20% of bins have $E_j \leq 5$, none $E_j \leq 1$ ('size restriction').

We obtain then $X^2 = 17.9$ for the Weibull model, and $X^2 = 92.2$ for the exponential distribution. The latter produces a p-value on the order of 10^{-18}, but the former has a p-value around 0.0013. Thus, while the data could not possibly have come from an exponential distribution, or anything like it, the Weibull distribution, while unlikely to have produced exactly these data, is a plausible candidate.

Age	Observed	Expected Exponential	Expected Weibull
0–4	8	22.7	5.4
5–9	13	21.5	19.3
10–14	22	15.7	25.3
15–19	39	11.5	22.7
20–24	25	8.4	15.7
25+	15	23.1	14.6

Table 1.2: χ^2 computation for fitting tyrannosaur data.

1.3 Overview of the course

Why do we need special statistical methods for lifetime data? Some reasons are:

- **Large samples** Other models, such as single-decrement models with time-varying transition rates, may be closer to the truth. We may have more elaborate multivariate parametric models for the transition rates, but they are unlikely to be precisely true. The problem then is that the parametric families will eventually be rejected, once the sample size is large enough — and since we may be concerned with statistical surveys of, for example, the entire population of the UK, the sample sizes will be very large indeed. Nonparametric or semiparametric methods will be better able to let the data speak for themselves.

- **Small samples** While nonparametric models allow the data to speak for themselves, sometimes we would prefer that they be somewhat muffled. When the number of observed deaths is small — which can be the case, even in a very large data set, when considering advanced ages, above 90, and certainly above 100, because of the small number of individuals who survive to be at risk, but also in children, because of the very low mortality rate — the estimates are less reliable, being subject to substantial random noise. Also, the mortality pattern changes over time, and we are often interested in future mortality, but only have historical data. A non-parametric estimate that precisely reflects the data at hand may reflect less well the underlying processes, and be ill-suited to projection into the future. Graduation (smoothing) and extrapolation methods have been developed to address these issues.

- **Incomplete observations** Some observations will be incomplete. We may not know the exact time of a death, but only that it occurred before a given time, or after a given time, or between two known times, a phenomenon called "censoring". (When we are informed only of the year of a death, but not the day or time, this is a kind of censoring. Or we may have observed only a sample of the population, with the sample being not entirely random, but chosen according to being alive at a certain date, or having died before a certain date, a phenomenon known as "truncation". We need special techniques to make use of these partial observations.) Since we are observing times, subjects who break off a study midway through provide partial information in a clearly structured way.

- **Successive events** A key fact about time is its sequence. A patient is infected, develops symptoms, has a diagnosis, a treatment, is cured or relapses, at some point dies. Some or all of these events may be considered as a progression, and we may want to model the

sequence of random times. Some care is needed to carry out joint maximum likelihood estimation of all transition rates in the model, from one or several individuals observed. This can be combined with time-varying transition rates.

- **Comparing lifetime distributions** We may wish to compare the lifetime distributions of different groups (e.g., smokers and nonsmokers; those receiving a traditional cholesterol medication and those receiving the new drug) or the effect of a continuous parameter (e.g., weight) on the lifetime distribution.

- **Changing rates** Mortality rates are not static in time, creating disjunction between period measures — looking at a cross-section of the population by age as it exists at a given time — and cohort measures — looking at a group of individuals born at a given time, and following them through life.

Lecture 2

Lifetime distributions

All the stochastic models in this course will be within the class of discrete state-space Markov processes which may be time inhomogeneous. We will not be using the general form of these models, but will be simplifying and specialising them substantially. What unifies this course is the nature of the questions we will be asking. In the standard theory of Markov processes, we focus early on stationary processes. Our models will not be stationary, because they have absorbing states. The key questions will concern the absorbing states: When the process is absorbed (the "lifetime"), and, in some models, which state absorbs it.

We need to be careful to distinguish between representations of the population and representations of the individual. In the present context, the Markov process always represents an individual. The population consists of some number of independently running copies of the basic Markov process. In simple cases — for instance, exponential mortality — the population-level process (total population at time t) will also be a Markov process, a "pure-death" chain. This raises the complication that there are usually two different kinds of time running: The "internal" time of the individual process, which usually represents age in some way, and calendar time. The full implications of these interacting time-frames — also called the cohort and the period perspective — are a major topic in demography, and we will only touch on them in this course.

2.1 Survival function and hazard rate (force of mortality)

As discussed in chapter 1, the simplest lifetime model is the single-decrement model: The individual is alive for some length of time L, at the end of which he/she becomes dead. This is a homogeneous Markov process if and only if L has an exponential distribution. In general, we may describe a lifetime distribution — which is simply the distribution of a nonnegative random variable — in several different ways:

cdf	$F(t) = \mathbb{P}\{L \leq t\};$
survival function	$S(t) = \bar{F}(t) = 1 - F(t) = \mathbb{P}\{L > t\};$
density function	$f(t) = dF/dt;$
hazard rate	$\lambda(t) = f(t)/\bar{F}(t)$

The hazard rate is also called mortality rate in survival contexts. The traditional name in demography is force of mortality. This may be thought of as the instantaneous rate of dying per unit time, conditioned on having already survived.s The *exponential distribution* with parameter $\lambda \in (0, \infty)$ is given by

cdf	$F(t) = 1 - e^{-\lambda t}$;
survival function	$\bar{F}(t) = e^{-\lambda t}$;
density function	$f(t) = \lambda e^{-\lambda t}$;
hazard rate	$\lambda(t) = \lambda$.

Thus, the exponential is the distribution with **constant** force of mortality, which is a formal statement of the "memoryless" property.

2.2 Residual lifetimes

Assume that there is an overall lifetime distribution, and every individual born has a random lifetime according to this distribution. Then, if we observe sombody now aged x, and we denote his residual lifetime $T - x$ by T_x, then we have

$$\bar{F}_{T_x}(t) = \bar{F}_{T-x|T>x}(t) = \frac{\bar{F}_T(x+t)}{\bar{F}_T(x)}, \qquad f_{T_x}(t) = f_{T-x|T>x}(t) = \frac{f_T(x+t)}{\bar{F}_T(x)}, \qquad t \geq 0. \tag{2.1}$$

So, any distribution of a full lifetime T is naturally associated with a family of conditional distributions of T given $T > x$.

2.3 Force of mortality

We now look more closely at the hazard rate, which may be defined as

$$h_T(t) = \mu_t = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \mathbb{P}(T \leq t + \varepsilon | T > t) = \lim_{\varepsilon \downarrow 0} \frac{\frac{1}{\varepsilon}\mathbb{P}(t < T \leq t + \varepsilon)}{\mathbb{P}(T > t)} = \frac{f_T(t)}{\bar{F}_T(t)}. \tag{2.2}$$

The density $f_T(t)$ is the (unconditional) infinitesimal probability to die at age t. The hazard rate $h_T(t)$ is the (conditional) infinitesimal probability to die at age t of an individual known to be alive at age t. It may seem that the hazard rate is a more complicated quantity than the density, but it is very well suited to modelling mortality. Whereas the density has to integrate to one and the distribution function (survival function) has boundary values 0 and 1, the force of mortality has no constraints, other than being nonnegative — though if "death" is certain the force of mortality has to integrate to infinity. Also, we can read its definition as a differential equation and solve

$$\bar{F}_T'(t) = -\mu_t \bar{F}_T(t), \quad \bar{F}(0) = 1 \qquad \Rightarrow \qquad \bar{F}_T(t) = \exp\left\{-\int_0^t \mu_s ds\right\}, \qquad t \geq 0. \tag{2.3}$$

We can now express the distribution of T_x as

$$\bar{F}_{T_x}(t) = \frac{\bar{F}_T(x+t)}{\bar{F}_T(x)} = \exp\left\{-\int_x^{x+t} \mu_s ds\right\} = \exp\left\{-\int_0^t \mu_{x+r} dr\right\}, \qquad t \geq 0. \tag{2.4}$$

Note that this implies that $h_{T_x}(t) = h_T(x+t)$, so it is really associated with age $x+t$ only, not with initial age x nor with time t after initial age. Also note that, given a measurable function $\mu : [0, \infty) \to \mathbb{R}$, $\bar{F}_{T_x}(0) = 1$ always holds, $\bar{F}_{T_x}$ decreasing if and only if $\mu \geq 0$. $\bar{F}_{T_x}(\infty) = 0$ if and only if $\int_0^\infty \mu_t dt = \infty$. This leaves a lot of modelling freedom via the force of mortality.

Densities can now be obtained from the definition of the force of mortality (and consistency) as $f_{T_x}(t) = \mu_{t+x} \bar{F}_{T_x}(t)$.

2.4 Defining mortality laws from hazards

We are now in the position to model mortality laws via their force of mortality. Clearly, the $\text{Exp}(\lambda)$ distribution has a constant hazard rate $\mu_t \equiv \lambda$, and the uniform distribution on $[0, \omega]$ has a hazard rate

$$h_T(t) = \frac{1}{\omega - t}, \qquad 0 \leq t < \omega. \tag{2.5}$$

Note that here $\int_0^\omega h_T(t)dt = \infty$ squares with $\bar{F}_T(\omega) = 0$ and forces the maximal age ω. This is a general phenomenon: distributions with compact support have a divergent force of mortality at the supremum of their support, and the singularity is not integrable.

The *Gompertz distribution* is given by $\mu_t = Be^{\theta t}$. More generally, *Makeham's law* is given by

$$\mu_t = A + Be^{\theta t}, \qquad \bar{F}_{T_x}(t) = \exp\left\{-At - m\left(e^{\theta(x+t)} - e^{\theta x}\right)\right\}, \qquad x \geq 0, t \geq 0, \tag{2.6}$$

for parameters $A > 0$, $B > 0$, $\theta > 0$; $m = B/\theta$. Note that mortality grows exponentially. If θ is big enough, the effect is very close to introducing a maximal age ω, as the survival probabilities decrease very quickly. There are other parameterisations for this family of distributions. The Gompertz distribution is named for British actuary Benjamin Gompertz, who in 1825 first published his discovery [Gom25] that human mortality rates over the middle part of life seemed to double at constant age intervals. It is unusual, among empirical discoveries, for having been confirmed rather than refuted as data have improved and conditions changed, and it (or Makeham's modification) serves as a standard model for mortality rates not only in humans, but in a wide variety of organisms. As an example, see Figure 2.1, which shows Canadian mortality rates from life tables produced by Statistics Canada (available at `http://www.statcan.ca:80/english/freepub/84-537-XIE/tables.htm`). Notice how close to a perfect line the mid-life mortality rates for both males and females is, when plotted on a logarithmic scale, showing that the Gompertz model is a very good fit.

Figure 2.1(b) shows the corresponding survival curves. It is worth recognising how much more informative the mortality rates are. in Figure 2.1(a) we see that male mortality is regularly higher than female mortality at all ages (and by a fairly constant ratio), we see several phases of mortality — early decline, jump in adolescence, then steady increase through midlife, and deceleration in extreme old age — whereas Figure 2.1(b) shows us only that mortality is accelerating overall, and that males have accumulated higher mortality by late life.

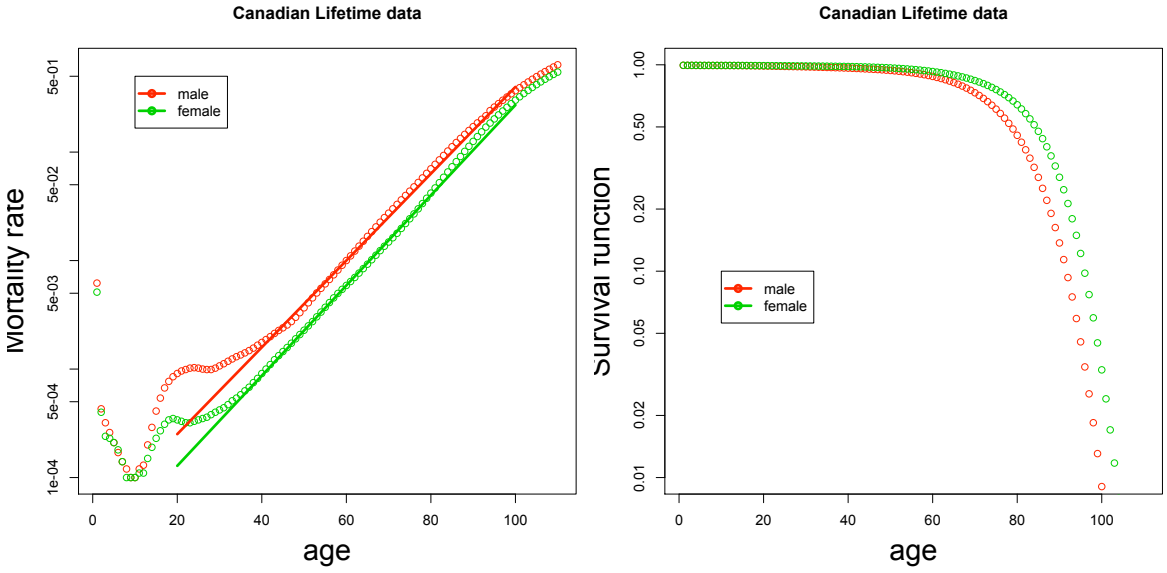
The *Weibull distribution* suggests a polynomial rather than exponential growth of mortality

$$\mu_t = kt^n, \qquad \bar{F}_{T_x}(t) = \exp\left\{-\frac{k}{n+1}\left((x+t)^{n+1} - x^{n+1}\right)\right\}, \qquad x \geq 0, t \geq 0, \tag{2.7}$$

for rate parameter $k > 0$ and exponent $n > 0$. The Weibull model is commonly used in engineering contexts to represent the failure-time distribution for machines. The Weibull distribution arises naturally as the lifespan of a machine with n redundant components, each of which has constant failure rate, such that the machine fails only when all components have failed. Later in the course we will discuss how to fit Weibull and Gompertz models to data.

Another class of distributions is obtained by replacing the parameter λ in the exponential distribution by a (discrete or continuous) random variable M. Then the specification of exponential conditional densities

$$f_{T|M=\lambda}(t) = \lambda e^{-\lambda t} \tag{2.8}$$



(a) Mortality rates

(b) Survival function

Figure 2.1: Canadian mortality data, 1995–7.

determines the unconditional density of T as

$$f_T(t) = \int_0^\infty f_{T,M}(t,\lambda)d\lambda = \int_0^\infty \lambda e^{-\lambda t} f_M(\lambda)d\lambda \qquad \text{or} \qquad f_T(t) = \sum_{\lambda>0} \lambda e^{-\lambda t}\mathbb{P}(M=\lambda). \quad (2.9)$$

Various special cases of exponential mixtures and other extensions of the exponential distribution have been suggested in a life insurance context. Some of these will be presented later.

E.g., for $M \sim \text{Geom}(p)$, i.e. $\mathbb{P}(M=k) = p^{k-1}(1-p)$, $k \geq 1$, we obtain

$$\begin{aligned}\bar{F}_T(t) &= \int_t^\infty f_T(s)ds = \int_t^\infty \sum_{k=1}^\infty f_{T|M=k}(s)p^{k-1}(1-p)ds \\ &= \sum_{k=1}^\infty \int_t^\infty ke^{-kt}p^{k-1}(1-p)ds = \frac{(1-p)e^{-t}}{1-pe^{-t}}\end{aligned}$$

and one easily deduces

$$f_T(t) = \frac{(1-p)e^{-t}}{(1-pe^{-t})^2}, \qquad t \geq 0.$$

The corresponding hazard rate is

$$h_T(t) = \frac{f_T(t)}{\bar{F}_T(t)} = \frac{1}{1-pe^{-t}},$$

which is an increasing but bounded.

2.5 Curtate lifespan

We have implicitly assumed that the lifetime distribution is continuous. However, we can always pass from a continuous random variable T on $[0, \infty)$ to a discrete random variable $K = [T]$, its integer part, on $\mathbb{N}$. If T models a lifetime, then K is called the associated *curtate lifetime*.

2.6 Single decrement model

The exponential model may also be represented as a Markov process. Let $\mathbb{S} = \{0, 1\}$ be our state space, with interpretation 0='alive' and 1='dead', and consider the Q-matrix

$$Q = \begin{pmatrix} -\mu & \mu \\ 0 & 0 \end{pmatrix}. \tag{2.10}$$

Then a continuous-time Markov chain $X = (X_t)_{t \geq 0}$ with $X_0 = 0$ and Q-matrix Q will have a holding time $T \sim \exp(\mu)$ in state 0 before a transition to 1, where it is absorbed, i.e.

$$X_t = \begin{cases} 0 & \text{if } 0 \leq t < T \\ 1 & \text{if } t \geq T \end{cases}. \tag{2.11}$$

The transition matrix is

$$P_t = e^{tQ} = \begin{pmatrix} e^{-\mu t} & 1 - e^{-\mu t} \\ 0 & 1 \end{pmatrix}.$$

It seems that this is an overly elaborate description of a simple model (diagrammed in Figure 2.2), but this viewpoint will be useful for generalisations. Also, the 'rate parameter' μ has a more concrete meaning, and the lack of memory property of the exponential distribution is also reflected in the Markov property: given that the chain is still in state 0 at time t (i.e. given $T > t$), the residual holding time (i.e. $T - t$) has conditional distribution $\text{Exp}(\mu)$.

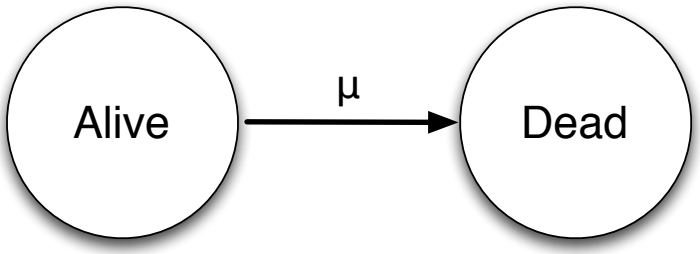


Figure 2.2: The single-decrement model.

This model may be generalised by allowing the transition rate μ to become an age-dependent rate function $t \mapsto \mu(t)$. This may be seen as a very special kind of inhomogeneous Markov process, or as a special kind of renewal process (one with only one transition). The general two-state model with transient state 'alive' and absorbing state 'dead', is called the 'single-decrement model'.

2.7 Mortality laws: Simple or Complex? Parametric or Nonparametric?

Consider the data for *Albertosaurus sarcophagus* in Table 1.1. We see here the estimated ages at death for 22 members of this species. Let us assume, for the sake of discussion, that these estimates are correct, and that our skeleton collection represents a simple random sample of all Albertosaurs that ever lived. If we assume that there was a large population of these dinosaurs, and that they died independently (and not, say, in a Cretaceous suicide pact), then these are 22 independent samples $T_1, \ldots, T_{22}$ of a random variable T whose distribution we would like to know. Consider the probabilities

$$q_x := \mathbb{P}\big\{x \leq T < x+1\big\}.$$

Then the number of individuals observed to have curtate lifespan x has binomial distribution $Bin(22, q_x)$. The MLE for a binomial probability is just the naïve estimate $\hat{q}_x =$ # successes/# trials (where a "success", in this case, is a death in the age interval under consideration). To compute $\hat{q}_2$, then, we observe that there were 22 Albertosaurs from our sample still alive on their 22 birthdays, of which one unfortunate met its maker in the following year: $\hat{q}_2 = 1/22 \approx 0.046$. As for $\hat{q}_3$, on the other hand, there were 21 Albertosaurs observed alive on their third birthdays, and all of them arrived safely at their fourth, making $\hat{q}_3 = 0/21$. This leads us to the peculiar conclusion that our best estimate for the probability of an albertosaur dying in its third year is 0.046, but that the probability drops to 0 in its fourth year, then becomes nonzero again in the fifth year, and so on. This violates our intuition that mortality rates should be fairly smooth as a function of age. This problem becomes even more extreme when we consider continuous lifetime models. With no constraints, the optimal estimator for the mortality distribution would put all the mass on just those moments when deaths were observed in the sample, and no mass elsewhere — in other words, infinite hazard rate at a finite set of points at which deaths have been observed, and 0 everywhere else.

As we see from Figure 1.1, the mortality distribution for the tyrannosaurs becomes much smoother and less erratic when we use larger bins for the histogram. This is no surprise, since we are then sampling from a larger baseline, leading to less random fluctuation. The simplest way to impose our intuition of regularity upon the estimators is to increase the time-step and reduce the number of parameters to estimate. An extreme version of this, of course, is to impose a parametric model with a small number of parameters. This is part of the standard tradeoff in statistics: a free, nonparametric model is sensitive to random fluctuations, but constraining the model imposes preconceived notions onto the data.

Notation: When the hazard rate μ_x is being assumed constant over each year of life, the continuous mortality rate has been reduced to a discrete set of parameters. What do we call these parameters? By convention, the value of μ that is in effect for all ages in $[x, x+1)$ is identified with just one age, namely $\mu_{x+\frac{1}{2}}$.

Lecture 3

Life Tables

Reading: Gerber Sections 2.4-2.5, CT4 Units 5-2, 6, 10-1
Further reading: Cox-Oakes Sections 4.1-4.4, Gerber Sections 11.1-11.5

Life tables represent a discretised form of the hazard function for a population, often together with raw mortality data. Apart from an aggregate table subsuming the whole population (of the UK, say), such tables exist for various groups of people characterized by their sex, smoking habits, job type, insurance level etc. This immediately raises interesting questions concerning the interdependence of such tables, but we focus here on some fundamental issues, which are already present for the single aggregate table.

We begin with a naïve, empirical approach. In Table 3.2 we see a life table for men in the UK, in the years 1990–2, as provided by the Office of National Statistics. In the column labelled E_x we see the number of years "exposed to risk" in age-class x. Since everyone alive is at risk of dying, this should be exactly the sum of the number of individuals alive in the age class in years 1990, 1991, and 1992. The 1991 number is obtained from the census of that year, and the other two years are estimated. The column d_x shows the number of men of the given age known to have died during this three-year period. The final column is $m_x := d_x/E_x$.

Again, this is an empirical fact, but we find ourselves in a quandary when we try to interpret it. What is m_x? If the number of deaths is reasonably stable from year to year, then m_x should be close to the fraction of men aged x who died each year. How close? The number of men at risk changes constantly, with each birthday, each death, each immigration or emigration. We sense intuitively that the effect of these changes would be small, but how small? And what would we do to compensate for this in a smaller population, where the effects are not negligible? How do we make projections about future states of the population?

AGE x	E_x	d_x	$m_x \times 10^5$	AGE x	E_x	d_x	$m_x \times 10^5$
0	1066867	8779	823	52	827414	4781	578
1	1059343	661	62	53	822603	5324	647
2	1054256	403	38	54	810731	5723	706
3	1047298	319	30	55	794930	6411	806
4	1037973	251	24	56	775350	6925	893
5	1022032	229	22	57	759747	7592	999
6	1003486	201	20	58	755475	8477	1122
7	989008	186	19	59	761913	9484	1245
8	976049	180	18	60	764497	10735	1404
9	981422	180	18	61	753706	11880	1576
10	988020	179	18	62	736868	12871	1747
11	984778	179	18	63	725679	14463	1993
12	950853	185	19	64	721743	16094	2230
13	909437	212	23	65	713576	17704	2481
14	891556	259	29	66	700666	19097	2726
15	913423	366	40	67	681977	20930	3069
16	954339	496	52	68	676972	22507	3325
17	1002077	758	76	69	678157	25127	3705
18	1057508	922	87	70	684764	27159	3966
19	1124668	930	83	71	600343	26508	4415
20	1163581	979	84	72	504808	24443	4842
21	1195366	1030	86	73	422817	22792	5391
22	1210521	1073	89	74	422480	24921	5899
23	1238979	1105	89	75	431321	27286	6326
24	1263313	1083	86	76	422822	29712	7027
25	1296300	1068	82	77	399257	30856	7728
26	1313794	1145	87	78	365168	30744	8419
27	1311662	1090	83	79	328386	30334	9237
28	1291017	1110	86	80	293014	29788	10166
29	1259644	1129	90	81	260517	28483	10933
30	1219278	1101	90	82	229149	27399	11957
31	1176120	1144	97	83	197322	25697	13023
32	1135091	1128	99	84	165896	23717	14296
33	1103162	1095	99	85	136103	20930	15378
34	1071474	1142	107	86	110565	18689	16903
35	1035587	1218	118	87	87989	16370	18605
36	1017422	1291	127	88	68443	13571	19828
37	1010544	1399	138	89	52151	11284	21637
38	1006929	1536	153	90	40257	9061	22508
39	1006500	1660	165	91	29000	7032	24248
40	1016727	1662	163	92	20124	5405	26858
41	1046632	1967	188	93	13406	4057	30263
42	1092927	2240	205	94	9392	3069	32677
43	1167798	2543	218	95	6446	2219	34424
44	1134652	2656	234	96	4384	1578	35995
45	1071729	2836	265	97	2795	1091	39034
46	974301	2930	301	98	1761	701	39807
47	955329	3251	340	99	1059	489	46176
48	914107	3354	367	100	624	292	46795
49	848419	3486	411	101	359	178	49582
50	815653	3836	470	102	216	118	54630
51	811134	4251	524	103	107	63	58879

Table 3.1: Male mortality data for England and Wales, 1990–2. From [Fox97] (available online at http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=333).

AGE x	ℓ_x	q_x	e_x	AGE x	ℓ_x	q_x	e_x
0	100000	0.0082	73.4	52	92997	0.0058	24.4
1	99180	0.0006	73.0	53	92461	0.0065	23.6
2	99119	0.0004	72.1	54	91865	0.0070	22.7
3	99081	0.0003	71.1	55	91219	0.0080	21.9
4	99052	0.0002	70.1	56	90486	0.0089	21.0
5	99028	0.0002	69.1	57	89682	0.0099	20.2
6	99006	0.0002	68.2	58	88791	0.0112	19.4
7	98986	0.0002	67.2	59	87800	0.0124	18.6
8	98967	0.0002	66.2	60	86714	0.0139	17.9
9	98950	0.0002	65.2	61	85505	0.0156	17.1
10	98932	0.0002	64.2	62	84168	0.0173	16.4
11	98914	0.0002	63.2	63	82710	0.0197	15.6
12	98896	0.0002	62.2	64	81078	0.0221	14.9
13	98877	0.0002	61.2	65	79290	0.0245	14.3
14	98855	0.0003	60.2	66	77347	0.0269	13.6
15	98826	0.0004	59.3	67	75267	0.0302	13.0
16	98786	0.0005	58.3	68	72992	0.0327	12.4
17	98735	0.0008	57.3	69	70605	0.0364	11.8
18	98660	0.0009	56.4	70	68037	0.0389	11.2
19	98574	0.0008	55.4	71	65391	0.0432	10.6
20	98492	0.0008	54.5	72	62567	0.0473	10.1
21	98410	0.0009	53.5	73	59610	0.0525	9.6
22	98325	0.0009	52.5	74	56481	0.0573	9.1
23	98238	0.0009	51.6	75	53246	0.0613	8.6
24	98150	0.0009	50.6	76	49982	0.0679	8.1
25	98066	0.0008	49.7	77	46590	0.0744	7.7
26	97986	0.0009	48.7	78	43125	0.0807	7.2
27	97900	0.0008	47.8	79	39643	0.0882	6.8
28	97819	0.0009	46.8	80	36145	0.0967	6.5
29	97735	0.0009	45.8	81	32651	0.1036	6.1
30	97647	0.0009	44.9	82	29270	0.1127	5.7
31	97559	0.0010	43.9	83	25971	0.1221	5.4
32	97465	0.0010	43.0	84	22800	0.1332	5.1
33	97368	0.0010	42.0	85	19763	0.1425	4.8
34	97272	0.0011	41.0	86	16946	0.1555	4.5
35	97168	0.0012	40.1	87	14310	0.1698	4.2
36	97053	0.0013	39.1	88	11881	0.1799	4.0
37	96930	0.0014	38.2	89	9744	0.1946	3.8
38	96796	0.0015	37.2	90	7848	0.2016	3.6
39	96648	0.0016	36.3	91	6266	0.2153	3.3
40	96489	0.0016	35.4	92	4917	0.2355	3.1
41	96332	0.0019	34.4	93	3759	0.2611	2.9
42	96151	0.0021	33.5	94	2777	0.2787	2.7
43	95954	0.0022	32.5	95	2003	0.2912	2.6
44	95745	0.0023	31.6	96	1420	0.3023	2.5
45	95521	0.0027	30.7	97	991	0.3232	2.3
46	95269	0.0030	29.8	98	670	0.3284	2.2
47	94982	0.0034	28.9	99	450	0.3698	2.1
48	94660	0.0037	28.0	100	284	0.3737	2.0
49	94313	0.0041	27.1	101	178	0.3909	1.9
50	93926	0.0047	26.2	102	108	0.4209	1.8
51	93486	0.0052	25.3	103	63	0.4450	1.7

Table 3.2: Life table for English men, computed from data in Table 3.1

3.1 Notation for life tables

q_x	Probability that individual aged x dies before reaching age $x+1$
p_x	Probability that individual aged x survives to age $x+1$
${}_tq_x$	Probability that individual aged x dies before reaching age $x+t$
${}_tp_x$	Probability that individual aged x survives to age $x+t$
l_x	Number of people who survive to age x. Note: This is based on starting with a fixed number l_0 of lives, called the Radix; most commonly, for human populations the radix is 100,000
d_x	Number of individuals who die aged x (from the standard population)
${}_tm_x$	Mortality rate between exact age x and exact age $x+t$
e_x	Remaining life expectancy at age x

Note the following relationships:

$$d_x = l_x - l_{x+1};$$
$$l_{x+1} = l_x p_x = l_x(1-q_x);$$
$${}_tp_x = \prod_{i=0}^{t-1} p_{x+i}$$

The quantities q_x may be thought of as the discrete analogue of the mortality rate — we will call it the *discrete mortality rate* or *discrete hazard function* — since it describes the probability of dying in the next unit of time, given survival up to age x. In Table 3.1 we show the life table computed from the raw data of Table 3.2. (It differs slightly from the official table, because the official table added some slight corrections. The differences are on the order of 1% in q_x, and much smaller in l_x.) The life table represents the effect of mortality on a nominal population starting with size l_0 called the Radix, and commonly fixed at 100,000 for large-population life tables. Imagine 100,000 identical individuals — a cohort — born on 1 January, 1900. In the column q_x we give the estimates for the probability of an individual who is alive on his x birthday dying in the next year, before his $x+1$ birthday. (We discuss these estimates later in the chapter.) Thus, we estimate that 820 of the 100,000 will die before their first birthday. The surviving $l_1 = 99,180$ on 1 January, 1901, face a mortality probability of 0.00062 in their next year, so that we expect 61 of them to die before their second birthday. Thus $l_2 = 99119$. And so it goes. The final column of this table, labelled e_x, gives remaining life expectancy; we will discuss this in section 4.2.

3.2 Continuous and discrete models

3.2.1 General considerations

The first decision that needs to be made in setting up a lifetime model is whether to model lifetimes as continuous or discrete random variables. On first consideration, the discrete approach may seem to recommend itself: after all, we are commonly concerned with mortality data given in whole years or, if not years, then whole numbers of months, weeks, or days. Real measurements are inevitably discrete multiples of some minimal unit of precision. In fact, though, discrete models for measured quantities are problematic because

- They tie the analysis to one unit of measurement. If you start by measuring lifespans in years, and restrict the model accordingly you have no way of even posing a question about, for instance, the effect of shifting the reporting date within the year.

- Discrete methods are comfortable only when the numbers are small, whereas moving down to the smallest measurable unit turns the measurements into large whole numbers. Once you start measuring an average human lifespan as 30000 days (more or less), real numbers become easier to work with, as integrals are easier than sums.

- It is relatively straightforward to embed discrete measures within a continuous-time model, by considering the integer part of the continuous random lifetime, called the curtate lifetime in actuarial terminology.

(Compare this to the suggestion once made by the physicist Enrico Fermi, that lecturers might take their listeners' investment of time more seriously if they thought of the 50-minute span of a lecture as a "microcentury".) The discrete model, it is pointed out by A. S. Macdonald in [Mac96] (and rewritten in [CT406, Unit 9]), "is not so easily generalised to settings with more than one decrement. Even the simplest case of two decrements gives rise to difficult problems," and involves the unnecessary complication of estimating an Initial Exposed To Risk. We will generally treat the continuous model as the fundamental object, and treat the discrete data as coarse representations of an underlying continuous lifetime. However, looking beyond the actuarial setting, there are models which really do not have an underlying continuous time parameter. For instance, in studies of human fertility, time is measured in menstrual cycles, and there simply are no intermediate chances to have the event occur.

3.2.2 Are life tables continuous or discrete?

The standard approach to life tables mixes the continuous and discrete, in sometimes confusing ways. The data upon which life tables are based are measured in discrete units, but in most applications we assume that the risk is actually continuous. If we were to observe a fixed number of individuals for exactly one year, and count the number of deaths at the end of the year, and if the number of deaths during the year were a small fraction of the total number at risk, it would hardly matter whether we chose a discrete or continuous model. As we discuss in chapter 5.3, the distinction becomes significant to the extent that the number of individuals at risk changes substantially over a single time unit; then we need to distinguish among Initial Exposed To Risk, Central Exposed To Risk, and the census approximation (see chapter 5.3).

The connection between discrete and continuous laws is fairly straightforward, at least in one direction. Suppose T is a lifetime with hazard rate μ_x at age x, and q_x is the probability of dying on or after birthday x, and before the $x+1$ birthday. Then

$$ {}_tq_x = e^{-\int_x^{x+t} \mu_s ds}. $$

Another way of putting this is to say that the discrete model may be embedded in the continuous model, by considering the discrete random variable $K = [T]$, called the associated curtate lifetime. The remainder (fractional part) $S = T - K = \{T\}$ can often be treated separately in a simplified way (see below). Clearly, the probability mass function of K on $\mathbb{N}$ is

given by

$$\mathbb{P}(K=n) = \mathbb{P}(n \le T < n+1) = \int_n^{n+1} f_T(t)dt = \bar{F}_T(n) - \bar{F}_T(n+1)$$
$$= \exp\left\{-\int_0^n \mu_T(t)dt\right\}\left(1 - \exp\left\{-\int_n^{n+1} \mu_T(t)dt\right\}\right)$$

and if we denote the one-year death probabilities (discrete hazard function) by

$$q_k = \mathbb{P}(K=k|K \ge k) = \frac{\mathbb{P}(K=k)}{\mathbb{P}(K \ge k)} = 1 - \exp\left\{-\int_k^{k+1} \mu_T(t)dt\right\}$$

and $p_k = 1 - q_k$, $k \in \mathbb{N}$, we obtain the probability of success after n independent Bernoulli trials with varying success probabilities q_k:

$$\mathbb{P}(K=n) = p_0 \dots p_{n-1} q_n.$$

Note that q_k only depends on the hazard rate between ages k and $k+1$. As a consequence, for $K_x = [T_x]$

$$\mathbb{P}(K_x = n) = p_x \dots p_{x+n-1} q_{x+n}$$

are also easily represented in terms of $(q_k)_{k\in\mathbb{N}}$.

3.3 Interpolation for non-integer ages

Suppose now that we have modeled the curtate lifetime K. The fractional part S of the lifetime is a random variable on the interval $[0,1]$, commonly modeled in one of the following ways:

Constant force of mortality $\mu(x)$ is constant on the interval $[k, k+1)$, and is called $\mu_T(k+\frac{1}{2})$, or sometimes $\mu_{k+\frac{1}{2}}$ when T is clear from the context. Then

$${}_1q_k = 1 - e^{-\mu_{k+\frac{1}{2}}}; \qquad \mu_{k+\frac{1}{2}} = -\ln p_k.$$

S has the distribution of an exponential random variable conditioned on $S < 1$, so it has density

$$f(s) = \mu_{k+\frac{1}{2}} \frac{e^{-\mu_{k+\frac{1}{2}} s}}{1 - e^{-\mu_{k+\frac{1}{2}}}}.$$

This assumption thus implies decreasing density of the lifetime through the interval. We also have, for $0 \le s \le 1$, and k an integer,

$${}_sp_k = \mathbb{P}(T > k+s|T>k) = \exp\left\{-\int_k^{k+s} \mu_t dt\right\} = \exp\left\{-s\mu_{k+\frac{1}{2}}\right\} = (1-q_k)^s.$$

Note that K and S are not independent, under this assumption.

Uniform If S is uniform on $[0,1)$, this implies that for $s \in [0,1)$,

$$f_T(k+s) = (\bar{F}_T(k) - \bar{F}_T(k+1)) = \bar{F}_T(k) q_k,$$
$${}_sq_k = s \cdot {}_1q_k,$$
$$\bar{F}_T(k+s) = \bar{F}_T(k)\big(1 - sq_k\big),$$
$$\mu_T(k+s) = \frac{f_T(k+s)}{\bar{F}_T(k+s)} = \frac{q_k}{1-sq_k}.$$

So this assumption implies that the force of mortality is increasing over the time unit. Note that μ is discontinuous at (some if not all) integer times unless $q_0 = \alpha = 1/n$ and $q_{x+1} = q_x/(1-q_x)$, i.e. $q_k = \frac{\alpha}{1-k\alpha}$, $k = 1, \ldots, n-1$, with $\omega = n$ maximal age. Usually, one accepts discontinuities.

Balducci ${}_{1-t}q_{k+t} = (1-t)q_k$ for $t \in [0,1)$, so that the probability of death in the remaining time $1-t$, having survived to $k+t$, is the product of the time left and the probability of death in $[k, k+1)$. There is a trivial identity that the probability of surving 1 time unit from time k is the probability of surviving t time units from time k, times the probability of surviving $1-t$ time units from time $k+t$. Thus

$$q_k = 1 - (1 - {}_tq_k) \cdot (1 - {}_{1-t}q_{k+t}) = 1 - (1 - {}_tq_k) \cdot (1 - (1-t)q_k),$$

so that

$${}_tq_k = 1 - \frac{1-q_k}{1-(1-t)q_k},$$

This implies that

$$\begin{aligned}
\bar{F}_T(k+t) &= \bar{F}_T(k)\mathbb{P}\big\{T > k+t \,\big|\, T > k\big\} = \bar{F}_T(k)\frac{1-q_k}{1-q_k+tq_k} \\
f_T(k+t) &= \frac{d}{dt}\bar{F}_T(k+t) = \frac{\bar{F}_T(k)q_k(1-q_k)}{(1-q_k+tq_k)^2}, \\
\mu_T(k+t) &= \frac{f_T(k+t)}{\bar{F}_T(k+t)} = \frac{q_k}{1-q_k+tq_k}
\end{aligned}$$

So this assumption implies that the force of mortality is decreasing over the time unit.

Once we have made one of these assumptions, we can reconstruct the full distribution of a lifetime T from the entries $(q_x)_{x\in\mathbb{N}}$ of a life table. When the force of mortality is small, these different assumptions are all equivalent to $\mu_{k+\frac{1}{2}} = q_k$. Notice again that the choice of a measurement unit for discretisation implies a certain level of smoothing, in continuous nonparametric life table computations. Taking the evidence at face value, we would have to say that we have observed zero mortality rate, except at the instants at which deaths were observed, where mortality jumps to ∞. Of course, we average over a period of time, either by imposing the constraint that mortality rates be step functions, constant over a single measurement unit (or multiple units, if we wish to impose additional smoothing, usually because the number of observations is small).

Moving in the other direction is not so straightforward. The continuous model cannot be embedded in the discrete model, for obvious reasons: within the framework of the discrete model, there is no such thing as a death midway through a time period. Traditionally, when the discrete nature of lifetable data has been in the foreground, a model of the fractional part, such as one of those listed above, has been adjoined to the model. As described in section 3.2.1, this approach quickly collapses under the weight of unnecessary complications, which is why we will always treat the continuous lifetime as the fundamental object, except when the lifetime truly is measured only in discrete units.

3.4 Crude estimation of life tables – discrete method

Since their invention in the 17th century, the basic methodology for life table has been to collect (from the church registry or whoever kept records of births and deaths) lifetimes, truncate to

integer lifetimes, count the numbers d_x of deaths between ages x and $x+1$, relate this to the numbers ℓ_x alive at age x, and use $\hat{q}_x^{(0)} = d_x/\ell_x$, or similar quantities as an estimate for the one-year death probability q_x.

In our model, the deaths are Bernoulli events with probability q_x, so we know that the Maximum Likelihood Estimator for q_x is $\hat{q}_x^{(0)} = \#\text{ successes}/\#\text{ trials} = d_x/\ell_x$ for $n = \ell_0$ independently observed curtate lifetimes $k_1, \dots, k_n$, observed from random variables with common probability mass function $(m(x))_{x\in\mathbb{N}}$ parameterized by $(q_x)_{x\in\mathbb{N}}$. If we denote $m(x) = (1-q_0)\dots(1-q_{x-1})q_x$, the likelihood is

$$\prod_{i=1}^{n} m(k^{(i)}) = \prod_{x\in\mathbb{N}} (m(x))^{d_x} = \prod_{x\in\mathbb{N}} (1-q_x)^{\ell_x - d_x} q_x^{d_x}, \tag{3.1}$$

where only $\max\{k_1,\dots,k_n\}+1$ factors in the infinite product differ from 1, and

$$d_x = d_x(k_1,\dots,k_n) = \#\left\{1 \le i \le n : k^{(i)} = x\right\},$$
$$\ell_x = \ell_x(k_1,\dots,k_n) = \#\left\{1 \le i \le n : k^{(i)} \ge x\right\}.$$

This product is maximized when its factors are maximal (the xth factor only depending on parameter q_x). An elementary differentiation shows that $q \mapsto (1-q)^{\ell-d}q^d$ is maximal for $\hat{q} = d/\ell$, so that

$$\hat{q}_x^{(0)} = \hat{q}_x^{(0)}(k_1,\dots,k_n) = \frac{d_x(k_1,\dots,k_n)}{\ell_x(k_1,\dots,k_n)}, \qquad 0 \le x \le \max\{k_1,\dots,k_n\}.$$

Note that for $x = \max\{k_1,\dots,k_n\}$, we have $\hat{q}_x^{(0)} = 1$, so no survival beyond the highest age observed is possible under the maximum likelihood parameters, so that $(\hat{q}^{(0)})_{0\le x\le \max\{k_1,\dots,k_n\}}$ specifies a unique distribution. (Varying the unspecified parameters q_x, $x > \max\{k_1,\dots,k_n\}$, has no effect.)

3.5 Crude life table estimation – continuous method

Alternatively, we can take a maximum likelihood approach on the continuous lifetimes, and obtain a different estimator. Assume that you observe $n = \ell_0$ independent lives $t_1,\dots,t_n$. Then the likelihood function is

$$\prod_{i=1}^{n} f_T(t_i) = \prod_{i=1}^{n} \mu_{t_i} \exp\left\{-\int_0^{t_i} \mu_s ds\right\} \tag{3.2}$$

Now assume that the force of mortality μ_s is constant on $[x, x+1)$, $x \in \mathbb{N}$ and denote these values by

$$\mu_{x+\frac{1}{2}} = -\ln(p_x) \qquad \left(\text{remember } p_x = \exp\left\{-\int_x^{x+1} \mu_s ds\right\}\right). \tag{3.3}$$

Then, the likelihood takes the form

$$\prod_{x\in\mathbb{N}} \mu_{x+\frac{1}{2}}^{d_x} \exp\left\{-\mu_{x+\frac{1}{2}}\tilde{\ell}_x\right\} \tag{3.4}$$

where only $\max\{t_1,\dots,t_n\}+1$ factors in the infinite product differ from 1, and

$$d_x = d_x(t_1,\dots,t_n) = \#\left\{1 \le i \le n : [t_i] = x\right\},$$
$$\tilde{\ell}_x = \tilde{\ell}_x(t_1,\dots,t_n) = \sum_{i=1}^{n} \int_x^{x+1} 1_{\{t_i > s\}} ds.$$

$\tilde{\ell}_x$ is called the *total exposed to risk*.

The quantities $\mu_{x+\frac{1}{2}}$, $x \in \mathbb{N}$, are the parameters, and we can maximise the product by maximising each of the factors. An elementary differentiation shows that $\mu \mapsto \mu^d e^{-\mu\ell}$ has a unique maximum at $\hat{\mu} = d/\ell$, so that

$$\hat{\mu}_{x+\frac{1}{2}} = \hat{\mu}_{x+\frac{1}{2}}(t_1,\dots,t_n) = \frac{d_x(t_1,\dots,t_n)}{\tilde{\ell}_x(t_1,\dots,t_n)}, \qquad 0 \le x \le \max\{t_1,\dots,t_n\}.$$

Since maximum likelihood estimators are invariant under reparameterisation (the range of the likelihood function remains the same, and the unique parameter where the maximum is obtained can be traced through the reparameterisation), we obtain

$$\hat{q}_x = \hat{q}_x(t_1,\dots,t_n) = 1 - \hat{p}_x = 1 - \exp\left\{-\hat{\mu}_{x+\frac{1}{2}}\right\} = 1 - \exp\left\{-\frac{d_x(t_1,\dots,t_n)}{\tilde{\ell}_x(t_1,\dots,t_n)}\right\}. \tag{3.5}$$

For small $d_x/\tilde{\ell}_x$, this is close to $d_x/\tilde{\ell}_x$, and therefore also close to d_x/ℓ_x.

Note that under $\hat{q}_x$, $x \in \mathbb{N}$, there is a positive survival probability beyond the highest observed age, and the maximum likelihood method does not fully specify a lifetime distribution, leaving free choice beyond the highest observed age.

3.6 Comparing continuous and discrete methods

There appears to be a contradiction between the discrete life-table estimation of section 3.4 and the continuous life-table estimation of section 3.5. While the models are different, there are questions to which both offer an answer, and the answers are different. In the discrete model, we estimate

$$\mathbb{P}\left\{T < x+1 \,\middle|\, T \ge x\right\} = q_x \approx \hat{q}_x = \frac{d_x}{\ell_x}.$$

The continuous model suggests that we estimate the same quantity by

$$\mathbb{P}\left\{T < x+1 \,\middle|\, T \ge x\right\} = 1 - e^{-\mu_{x+\frac{1}{2}}} \approx 1 - e^{-\hat{\mu}_{x+\frac{1}{2}}} = 1 - e^{-d_x/\tilde{\ell}_x} \le \frac{d_x}{\tilde{\ell}_x}. \tag{3.6}$$

If we take ℓ_x as a substitute for $\tilde{\ell}_x$, then, the continuous model gives a strictly smaller answer, unless $d_x = 0$. Why is that? The difference here is that the continuous model presumes that individuals are dying all through the year, making $\tilde{\ell}_x$ somewhat smaller than ℓ_x. In fact, if we make the estimate $\tilde{\ell}_x \approx \ell_x - d_x/2$ (so presuming that those who died lived on average half a year), substituting the Taylor series expansion into (3.6) shows that in the continuous model

$$\begin{aligned}\mathbb{P}\left\{T < x+1 \,\middle|\, T \ge x\right\} &= \frac{d_x}{\ell_x - d_x/2} - \frac{d_x}{2(\ell_x - d_x/2)^2} + o\left(\left(\frac{d_x}{\ell_x - d_x/2}\right)^3\right) \\ &= \frac{d_x}{\ell_x} + o\left(\left(\frac{d_x}{\ell_x - d_x/2}\right)^3\right).\end{aligned}$$

That is, when the mortality fraction d_x/ℓ_x is small, the estimates agree up to second order in d_x/ℓ_x.

3.7 An example: Fractional lifetimes can matter

Imagine an insurance company that insures valuable pieces of construction machinery, which we will call piddledonks. For safety reasons, piddledonks cannot be used more than 3 years, but they may fail before that time. The company has records on 1000 of these machines, summarised in Table 3.3. That is, 100 failed in their first year (age 0), 400 in the second year, and 400 in the third year of operation. The last column shows the estimated failure probabilities.

Table 3.3: Life table for piddledonks.

age x	l_x	d_x	q_x
0	1000	100	0.10
1	900	400	0.44
2	500	400	0.80

Suppose the company sells insurance policies that pay £1000 when a piddledonk fails. The fair price for such a contract will be £100 for a new-built piddledonk. (That is, the price equal to the expected value of the contract; obviously, a company that wants to cover its costs and even turn a profit needs to sell its insurance somewhat above the nominal fair price.) It will be £444 for a piddledonk on its first birthday, and £800 for a piddledonk on its second birthday. Suppose, though, someone comes with a piddledonk that is 18 months old, and wishes to buy insurance for the next half year. What would be the fair price?

We have no data on when in the year failure occurs. It is possible, in principle, that piddledonks fail only on their birthdays; if they survive that day, they're good for the rest of the year. In that case, the insurance could be free, since the probability of a failure in the second half year is 0. This seems implausible, though. Suppose we adopt the constant-hazard model. Calling the constant hazard μ, we see that $p_1 = e^{-\mu}$, and

$$p_1 = {}_{0.5}p_1 \cdot {}_{0.5}p_{1.5}. \tag{3.7}$$

Thus,

$${}_{0.5}p_{1.5} = {}_{0.5}p_1 = e^{-\mu/2} = \sqrt{p_1} = \sqrt{1-q_1} = \sqrt{.555} = 0.745,$$

and ${}_{0.5}q_{1.5} = 0.255$, and the fair price for the half year of insurance is £255. Suppose, on the other hand, we adopt the uniform model for S. We still have (3.7), but now

$${}_{0.5}p_1 = 1 - {}_{0.5}q_1 = 1 - \frac{1}{2}{}_1q_1,$$

so that

$${}_{0.5}p_{1.5} = \frac{p_1}{1 - \frac{1}{2}{}_1q_1} = \frac{0.555}{.778} = 0.713,$$

implying that the fair price for this insurance would be £287.

Lecture 4

Cohorts and Period Life Tables

4.1 Types of life tables

You may have noticed a logical fallacy in the arguments of sections 3.4 and 3.5. The life expectancy at birth should be the average length of life of individuals born in that year. Of course, we would have to go back to about 1890 to find a birth year whose cohort — the individuals born in that year — have completed their lives, so that the average lifespan can be computed as an average.

Consider, for instance, the discrete-time non-homogeneous model. "Time" in the model is individual age: An individual starts out at age 0, then progresses to age 1 if she survives, and so on. We estimate the probability of dying aged x by dividing the number of deaths observed age x by the number of individuals observed to have been at that age.

In our life-tables, called period life tables, these numbers came from a census of the individuals alive at one particular time, and the count of those who died in the same year, or period of a few years. No individual experiences those mortality rates. Those born in 2009 will experience the mortality rates for age 10 in 2019, and the mortality rates for age 80 in 2089. Putting together those mortality rates would give us a cohort life table. (Actually, this is not precisely true. You might think about why not. The answer is given in a footnote.[1]) If, as has been the case for the past 150 years, mortality rates decline in the interval, that means that the survival rates will be higher than we see in the period table.

We show in Figure 4.1 a picture of how a cohort life table for the 1890 cohort would be related to the sequence of period life tables from the 1890s through the 2000s. The mortality rates for ages 0 through 9 (thus ${}_1q_0$, ${}_4q_1$, ${}_5q_5$)[2] are on the 1890s period life table, while their mortality rates for ages 10 through 19 are on the 1900–1909 period life table, and so on. Note that the mortality rates for the 1890s period life table yield a life expectancy at birth $e_0 = 44.2$ years. That is the average length of life that babies born in those years would have had, if their mortality in each year of their lives had corresponded to the mortality rates which were realised in for the whole population in the year of their birth. Instead, though, those that survived their early years entered the period of late-life high mortality in the mid- to late 20th century, when mortality rates were much lower. It may seem surprising, then, that the life expectancy for the

[1] The main difference between a cohort life table and the life table constructed from the corresponding age classes of successive period life tables is immigration: The cohort life table for 1890 should include, in the row for (let us say) ages 60–4 the mortality rates of those born in 1890 in the relevant region — England and Wales in this case — who are still alive at age 60. But these are not identical to the 60 year old men living in England and Wales in 1950. Some of the original cohort have moved away, and some residing in the country were not born there.

[2] Actually, we have given μ_x for the intervals $[0, 1)$, $[1, 5)$, and $[5, 10)$. We compute ${}_1q_0 = 1-e^{-\mu_0}$, ${}_4q_1 = 1-e^{-4\mu_1}$, ${}_5q_5 = 1 - e^{-5\mu_5}$.

cohort life table only goes up to 44.7 years. Is it true that this cohort only gained 6 months of life on average, from all the medical and economic progress that took place during their lives?

Yes and no. If we look more carefully at the period and cohort life tables in Table 4.1 we see an interesting story. First of all, a substantial fraction of potential lifespan is lost in the first year, due to the 17% infant mortality, which is obviously the same for the cohort and period life tables. 25% died before age 5. If mortality to age 5 had been reduced to modern levels — close to zero — the period and cohort mortality would both be increased by about 14 years. Second, notice that the difference in life expectancies jumps to over 5 years at age 30. Why is that? For the 1890 cohort, age 30 was 1920 — after World War I, and after the flu pandemic. The male mortality rate in this age class was around 0.005 in 1900–9, and less than 0.004 in 1920–9. Averaged over the intervening decade, though, male mortality was close to 0.02. (Most of the effect is due to the war, as we see from the fact that it almost exclusively is seen in the male mortality; female mortality in the same period shows a slight tick upward, but it is on the order of 0.001.) One way of measuring the horrible cost of that war is to see that for the generation of men born in the 1890s, that was most directly affected, the advances of the 20th century procured them on average about 4 years of additional life, relative to what might have been expected from the mortality rates in the year of their birth. Of these 4 years, $3\frac{1}{2}$ were lost in the war. Another way of putting this is to see that the approximately 4.5 million boys born in the UK between 1885 and 1895 lost cumulatively about 16 million years of potential life in the war.

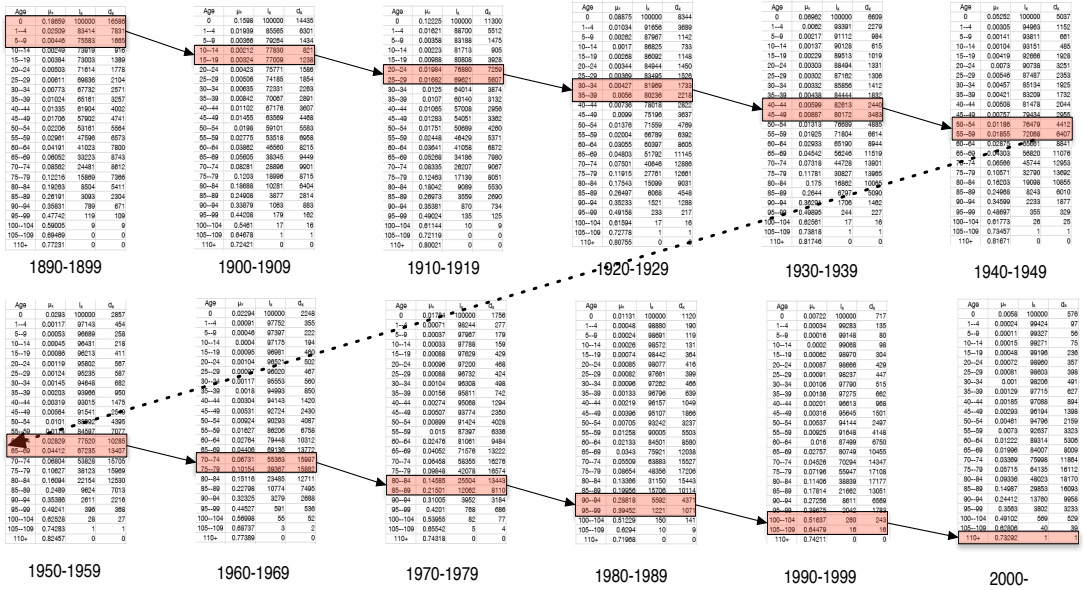


Figure 4.1: Decade period life tables, with the pieces joined that would make up a cohort life table for individuals born in 1890.

There are, in a sense, three basic kinds of life tables:

1. Cohort life table describing a real population. These make most sense in a biological

x	μ_x	ℓ_x	d_x	e_x
0	0.187	100000	17022	44.2
1	0.025	82978	7923	51.7
5	0.004	75055	1655	52.8
10	0.002	73400	908	48.9
15	0.004	72492	1379	44.5
20	0.005	71113	1766	40.3
25	0.006	69347	2087	36.2
30	0.008	67260	2550	32.2
35	0.010	64710	3229	28.4
40	0.013	61481	3970	24.7
45	0.017	57511	4703	21.2
50	0.022	52808	5515	17.8
55	0.030	47293	6508	14.6
60	0.042	40785	7710	11.6
65	0.061	33075	8636	8.9
70	0.086	24439	8511	6.5
75	0.122	15928	7281	4.5
80	0.193	8647	5346	2.7
85	0.262	3301	2410	1.7
90	0.358	891	742	0.9
95	0.477	149	135	0.5
100	0.590	14	13	0.3
105	0.695	1	1	0.2
110	0.772	0	0	0.0

(a) Period life table for men in England and Wales 1890–9

x	μ_x	ℓ_x	d_x	e_x
0	0.187	100000	17022	44.7
1	0.025	82978	7923	52.3
5	0.004	75055	1655	53.5
10	0.002	73400	774	49.6
15	0.003	72626	1167	45.1
20	0.020	71459	6749	40.8
25	0.017	64710	5219	39.7
30	0.004	59491	1257	37.9
35	0.006	58234	1608	33.6
40	0.006	56626	1671	29.5
45	0.009	54955	2384	25.3
50	0.012	52571	3027	21.3
55	0.019	49544	4388	17.5
60	0.028	45156	5956	14.0
65	0.044	39200	7760	10.8
70	0.067	31440	8985	8.0
75	0.102	22455	8940	5.7
80	0.146	13515	6997	3.8
85	0.215	6518	4294	2.3
90	0.288	2224	1697	1.4
95	0.395	527	454	0.8
100	0.516	73	67	0.4
105	0.645	6	6	0.2
110	0.733	0	0	0.0

(b) Cohort life table for the 1890 cohort of men in England and Wales

Table 4.1: Period and cohort tables for England and Wales. The period table is taken directly from the Human Mortality Database http://www.mortality.org/. The cohort table is taken from the period tables of the HMD, not copied from their cohort tables.

context, where there is a small and short-lived population. The ℓ_x numbers are actual counts of individuals alive at each time, and the rest of the table is simply calculated from these, giving an alternative descriptions of survival and mortality.

2. Period life tables, which describe a notional cohort (usually starting with radix ℓ_0 being a nice round number) that passes through its lifetime with mortality rates given by the q_x. These q_x are estimated from data such as those of Table 3.2, giving the number of individuals alive in the age class during the period (or number of years lived in the age class) and the number of deaths.

3. Synthetic cohort life tables. These take the q_x numbers from a real cohort, but express them in terms of survival ℓ_x starting from a rounded radix.

4.2 Life Expectancy

4.2.1 What is life expectancy?

One of the most interesting (and most discussed) features of life tables is the life expectancy. It has an intuitive meaning — the average length of life — and is commonly used as a summary of the life table, to compare mortality between countries, regions, and subpopulations. For instance, Table 4.2 shows the estimated life expectancy in some rich and poor countries, ranging from 37.2 years for a man in Angola, to 85.6 years for a woman in Japan. The UK is in between (though, of course, much closer to Japan), with 76.5 years for men and 81.6 years for women.

Table 4.2: 2009 Life expectancy at birth (LE) in years and infant mortality rate per thousand live births (IMR) in selected countries, by sex. Data from US Census Bureau. International Database available at http://www.census.gov/ipc/www/idb/idbprint.html

Country	IMR	IMR male	IMR female	LE	LE male	LE female
Angola	180	192	168	38.2	37.2	39.2
France	3.33	3.66	2.99	81.0	77.8	84.3
India	30.1	34.6	25.2	69.9	67.5	72.6
Japan	2.79	2.99	2.58	82.1	78.8	85.6
Russia	10.6	12.1	8.9	66.0	59.3	73.1
South Africa	44.4	48.7	40.1	49.0	49.8	48.1
United Kingdom	4.85	5.40	4.28	79.0	76.5	81.6
United States	6.26	6.94	5.55	78.1	75.7	80.7

Life expectancies can vary significantly, even within the same country. For example, the UK Office of National Statistics has published estimates of life expectancy for 432 local areas in the UK (available at http://www.statistics.gov.uk/life-expectancy/default.asp). We see there that, for the period 2005–7, men in Kensington and Chelsea had a life expectancy of 83.7 years, and women 87.8 years; whereas in Glasgow (the worst-performing area) the corresponding figures were 70.8 and 77.1 years. Overall, English men live 2.7 years longer on average than Scottish men, and English women 2.0 years longer.

When we think of lifetimes as random variables, the life expectancy is simply the mathematical expectation $\mathbb{E}[T]$. By definition,

$$\mathbb{E}[T] = \int_0^\infty x f_T(x) dx.$$

Integration by parts, using the fact that $f_T = -\bar{F}_T'$, turns this into a much more useful form,

$$\mathbb{E}[T] = -t\bar{F}_T(t)\Big|_0^\infty + \int_0^\infty \bar{F}_T(t)dt = \int_0^\infty \bar{F}_T(t)dt = \int_0^\infty e^{-\int_0^t \mu_s ds} dt. \tag{4.1}$$

That is, the life expectancy may be computed simply by integrating the survival function. The discrete form of this is

$$\mathbb{E}[K] = \sum_{k=0}^\infty k\mathbb{P}\{K = k\} = \sum_{k=0}^\infty \mathbb{P}\{K > k\}. \tag{4.2}$$

Applying this to life tables, we see that the expected curtate lifetime is

$$\mathbb{E}[K] = \sum_{k=0}^{\infty} \mathbb{P}\{K > k\} = \sum_{k=1}^{\infty} \frac{l_k}{l_0} = \sum_{k=1}^{\infty} p_0 \cdots p_{k-1}.$$

Note that expected future lifetimes can be expressed as

$$\mathring{e}_x := \mathbb{E}[T_x] = \int_x^{\infty} \exp\left\{-\int_x^t \mu_s ds\right\} dt \qquad \text{and} \qquad e_x := \mathbb{E}[K_x] = \sum_{k\in\mathbb{N}} p_x \dots p_{x+k} = \sum_{k=x+1}^{\omega} \frac{l_k}{l_x}.$$

We see that $e_x \leq \mathring{e}_x < e_x + 1$. For sufficiently smooth lifetime distributions, $\mathring{e}_x \approx e_x + \frac{1}{2}$ will be a good approximation.

For variances, formulas in terms of $y \mapsto \mu_y$ and $(p_x)_{x\geq 0}$ can be written down, but do not simplify as neatly. Also the approximation $Var(T_x) \approx Var(K_x) + \frac{1}{12}$ requires rougher arguments: this follows e.g. if we assume that $S_x = T_x - K_x$ is independent of K_x and uniformly distributed on $[0, 1]$.

4.2.2 Example

Table 4.3 shows a life table based on the mortality data for tyrannosaurs from Table 1.1. Notice that the life expectancy at birth $e_0 = 16.0$ years is exactly what we obtain by averaging all the ages at death in Table 4.3.

age	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
d_x	0	0	3	1	1	3	2	1	2	4	4	3	4	3	8
l_x	103	103	103	100	99	98	95	93	92	90	86	82	79	75	72
q_x	0.00	0.00	0.03	0.01	0.01	0.03	0.02	0.01	0.02	0.04	0.05	0.04	0.05	0.04	0.11
e_x	16.0	15.0	14.0	13.5	12.6	11.7	11.1	10.3	9.4	8.7	8.1	7.5	6.7	6.1	5.4

age	15	16	17	18	19	20	21	22	23	24	25	26	27	28
d_x	4	4	7	10	6	3	10	8	4	3	0	3	0	2
l_x	64	60	56	49	39	33	30	20	12	8	5	5	2	2
q_x	0.06	0.07	0.12	0.20	0.15	0.09	0.33	0.40	0.33	0.38	0.00	0.60	0.00	1.00
e_x	5.0	4.4	3.7	3.2	3.0	2.6	1.8	1.7	1.8	1.8	1.8	0.8	1.0	0.00

Table 4.3: Life table for tyrannosaurs, based on data from Table 1.1.

4.2.3 Life expectancy and mortality

The connection between life expectancy and mortality is somewhat subtle. It is well known that life expectancy at birth — e_0 — has been rising for well over a century. For males it is 73.4 years on the 1990-2 UK life table, but was only 44.1 years on the life table a century before. However, it would be a mistake to suppose this means that a typical man was dying at an age that we now consider active middle-age. This becomes clearer when we look at the remaining life expectancy at age 44. In 1990 it was 31.6 years; in 1890 it was 22.1 years. Less, to be sure, but still a substantial number of years remaining. The low average length of life in 1890 was determined in large part by the number of zeroes being included in the average.

Imagine a population in which everyone dies at exactly age 75. The expectation of life remaining at age x would then be exactly $75 - x$. While that is not, of course, our true situation, mortality in much of the developed world today is quite close to this extreme: There is almost no randomness, as witnessed by the fact that the remaining life expectancy column of the lifetable

marches monotonously down by one year per year lived. The only exception is at the beginning — the newborn has only lost about 0.4 remaining years for the year it has lived. This is because the mortality in the first year is fairly high, so that overcoming that hurdle gives a significant boost to ones remaining life expectancy. We can compute

$$e_0 = p_0(1 + e_1).$$

This follows either from (4.2), or directly from observing that if someone survives the first year (which happens with probability p_0) he will have lived one year, and have (on average) e_1 years remaining. Thus,

$$q_0 = 1 - \frac{e_0}{1 + e_1} = \frac{1 + e_1 - e_0}{1 + e_1} = \frac{.6}{74.4} = 0.008,$$

which is approximately right. On the 1890 life table we see that the life expectancy of a newborn was 44.1 years, but this rose to 52.2 years for a boy on his first birthday. This can only mean that a substantial portion of the children died in infancy. We compute the first-year mortality as $q_0 = (1 + 52.2 - 44.1)/53.2 = 0.17$, so about one in six.

How much would life expectancy have been increased simply by eliminating infant mortality — that is, mortality in the first year of life? In that case, all newborns would have reached their first birthday, at which point they would have had 52.2 years remaining on average — thus, 53.2 years in total. Today, with infant mortality almost eliminated, there is only a potential 0.6 years remaining to be achieved from further reductions.

4.3 An example of life-table computations

Suppose we are studying a population of creatures that live a maximum of 4 years. For simplicity, we will assume that births all occur on 1 January. (The complications of births going on throughout the year will be addressed in lecture 5.) The entire population is under observation, and all deaths are recorded. We make the following observations:

Year 1: 300 born, 100 die.

Year 2: 350 born, 150 die. 20 1-year-olds die.

Year 3: 400 born, 100 die. 40 1-year-olds die. 90 2-year-olds die.

Year 4: 300 born, 50 die. 75 1-year-olds die. 100 2-year-olds die. 90 3-year-olds die.

In Table 4.4 we compute different life tables from these data. The two cohort life tables (Tables 4.4(a) and 4.4(b)) are fairly straightforward: We start by writing down ℓ_0 (the number of births in that cohort) and then in the d_x column the number of deaths in each year from that cohort. Subtracting those successively from ℓ_0 yields the number of survivors in each age class ℓ_x, and $q_x = d_x/\ell_x$. Finally, we compute the remaining life expectancies:

$$\begin{aligned} e_0 &= \frac{\ell_1}{\ell_0} + \frac{\ell_2}{\ell_0} + \frac{\ell_3}{\ell_0} \\ e_1 &= \frac{\ell_2}{\ell_1} + \frac{\ell_3}{\ell_1} \\ e_2 &= \frac{\ell_3}{\ell_2}. \end{aligned}$$

x	d_x	ℓ_x	q_x	e_x
0	100	350	0.333	1.57
1	20	200	0.10	1.35
2	90	180	0.50	0.50
3	90	90	1.0	0

(a) Cohort 1 life table

x	d_x	ℓ_x	q_x	e_x
0	150	350	0.43	1.2
1	40	200	0.20	1.1
2	100	160	0.625	0.375
3	60	60	1.0	0

(b) Cohort 2 life table

x	q_x	ℓ_x	ℓ_x	e_x
0	0.167	1000	167	1.69
1	0.25	833	208	1.03
2	0.625	625	0.375	
3	1.0	235	1.0	0

(c) Period life table for year 4

Table 4.4: Alternative life tables from the same data.

The period life table is computed quite differently. We start with the q_x numbers, which come from different cohorts:

q_0 comes from cohort 4 newborn deaths;
q_1 comes from cohort 3 age 1 deaths;
q_2 comes from cohort 2 age 2 deaths;
q_3 comes from cohort 1 age 3 deaths.

We then write in the radix $\ell_0 = 1000$. Of 1000 individuals born, with $q_0 = 0.167$, we expect 167 to die, giving us our d_0. Subtracting that from ℓ_0 tells us that $\ell_1 = 833$ of the 1000 newborns live to their first birthday. And so it continues. The life expectancies are computed by the same formula as before, but now the interpretation is somewhat different. The cohort remaining life expectancies were the same as the actual average number of (whole) years remaining for the population of individuals from that cohort who reached the given age. The period remaining life expectancies are fictional, telling us how many individuals would have remained alive if we had a cohort of 1000 that experienced in each age the same mortality rates that were in effect for the population in year 4.

Lecture 5

Central exposed to risk and the census approximation

Reading: CT4 Units 6-2 and 10, Cox-Oakes Section 1.3, Gerber Section 11.1
Further reading: Cox-Oakes Chapter 3

5.1 Censoring

The term 'Censoring' refers to various types of incomplete information. The simplest example occurs in any study where processes (e.g. lifetimes in a single or multiple decrement framework) are observed over a limited time range, as is unavoidably imposed since the study cannot go on until all participants have died; if the end of the study is a predetermined fixed time, then the fact that a participant survives bears important information, so survivors must be given appropriate consideration in the likelihood.

In a single (or multiple) decrement model, this is e.g. the probability of survival: if r individuals are observed for t years (or prior death), then the likelihood contribution from those dying at time $s_i < t$, say, is the density $f_T(s_i)$; the likelihood contribution of those surviving to the end of the study at time t is the probability of survival $\bar{F}_T(t)$.

This is an example of *right censoring*, which, more generally, also occurs, when participants withdraw from the study for other exterior reasons (e.g. expiry date of an insurance policy). More general types of censoring will be treated later.

5.2 Insurance data

Insurance data have several special features. In the best of cases, we have full information from each person insured as follows; for a simple life assurance paying death benefits on death only, for individual m:

- date of birth b_m
- date of entry into observation: policy start date x_m
- reason for exit from observation (death $D_m = 1$, or expiry/withdrawal $D_m = 0$)
- date of exit from observation Y_m

This then easily translates into a likelihood

$$1_{\{D_m=1\}} f_{T_{x_m-b_m}}(Y_m - x_m) + 1_{\{D_m=0\}} \bar{F}_{T_{x_m-b_m}}(y_m - x_m) = \mu_{Y_m-b_m}^{D_m} \exp\left\{ - \int_{x_m-b_m}^{Y_m-b_m} \mu_t dt \right\}, \quad (5.1)$$

and it is clear how much time this individual was exposed to risk at age x, i.e. aged $[x, x+1)$ for all $x \in \mathbb{N}$. We can calculate the Central exposed to risk E_x^c as the aggregate quantity across all individuals exactly. We can also read off the number of deaths aged $[x, x+1)$, d_x, and hence

$$\hat{\mu}_{x+\frac{1}{2}} = \frac{d_x}{E_x^c} \quad (5.2)$$

This is the maximum likelihood estimator under the assumption of a constant force of mortality on $[x, x+1)$. Note that this estimator conforms with the Principle of Correspondence which states that

> A life alive at time t should be included in the exposure at age x at time t if and only if, were that life to die immediately, he or she would be counted in the death data d_x at age x.

In practice, data are often not provided in this form and approximations are required. E.g., policy start and end dates may not be available; instead, only total numbers of policies per age group at annual census dates are provided, and there is ambiguity as to when individuals change age group between the census dates. The solution to the problem is called the census approximation.

The key point is that we can tolerate a substantial amount of uncertainty in the numerator and the denominator (number of events and total time at risk), but failing to satisfy the Principle of Correspondence can be disastrous. For example, [ME05] analyses the "Hispanic Paradox," the observation that Latin American immigrants in the USA seem to have substantially lower mortality rates than the native population, despite being generally poorer (which is usually associated with shorter lifespans). This difference is particularly pronounced at more advanced ages. Part of the explanation seems to be return migration: Some old hispanics return to their home countries when they become chronically ill or disabled. Thus, there are some members of this group who count as part of the US hispanic population for most of their lives, but whose deaths are counted in their home-country statistics.

5.3 Census approximation

The task is to approximate E_x^c (and often also d_x) given census data. There are various forms of census data. The most common one is

$$P_{x,k} = \text{Number of policy holders aged } [x, x+1) \text{ at time } k = 0, \ldots, n.$$

The problem is that we do not know policy start and end dates. The basic *assumption* of the census approximation is that the number of policies changes linearly between any two consecutive census dates. It is easy to see that

$$E_x^c = \int_0^n P_{x,t} dt \quad (5.3)$$

We only know the integrand at integer times, and the linearity approximation gives

$$E_x^c \approx \sum_{k=1}^{n} \frac{1}{2}(P_{x,k-1} + P_{x,k}). \quad (5.4)$$

This allows us to estimate $\mu_{x+\frac{1}{2}}$ if we also know d_x, the number of deaths aged x.

Now assume that, in fact, you are not given d_x but only calender years of birth and death leading to

$$d'_x = \text{Number of deaths aged } x \text{ on the birthday in the calendar year of death.}$$

Then, some of the deaths counted in d'_x will be deaths aged $x-1$, not x, in fact we should view d'_x as containing deaths aged in the interval $(x-1, x+1)$, but not all of them. If we assume that birthdays are uniformly spread over the year, we can also specify that the proportion of deaths counted under d'_x changes linearly from 0 to 1 and back to 0 as $x-1$ increases to x and $x+1$.

In order to estimate a force of mortality, we need to identify the *corresponding* (approximation to) Central exposed to risk. The Principle of Correspondence requires

$$E^{c\prime}_x = \int_0^n P'_{x,t} dt, \tag{5.5}$$

where

$$P'_{x,t} = \text{Number of policy holders at } t \text{ with } x\text{th birthday in calendar year } [t].$$

Again, suppose we know the integrand at integer times. Here the linear approximation requires some care, since the policy holders do not change age group continuously, but only at census dates. Therefore, all continuing policy holders counted in $P'_{x,k-1}$ will be counted in $P'_{x,t}$ for all $k-1 \le t < k$, but then in $P'_{x+1,k}$ at the next census date. Therefore

$$E^{c\prime}_x \approx \sum_{k=1}^{n} \frac{1}{2}(P'_{x,k-1} + P'_{x+1,k}). \tag{5.6}$$

The ratio $d'_x/E^{c\prime}_x$ gives a slightly smoothed (because of the wider age interval) estimate of μ_x (and not $\mu_{x+\frac{1}{2}}$). Note however that it is *not* clear if this estimate is a maximum likelihood estimate for μ_x under any suitable model assumptions such as constancy of the force of mortality between half-integer ages.

Some other types of data appear on Assignment 3. The general problem is always to identify the corresponding central exposed to risk and what the ratio of death counts and its central exposed to risk estimates.

5.4 Lexis diagrams

A graphical tool that helps in making sense of estimates like the census approximation is the Lexis diagram.[1] These reduce the three dimensions of demographic data — date, age, and moment of birth — to two, by an ingenious application of the diagonal.

Consider the diagram in Figure 5.1. The horizontal axis represents calendar time (which we will take to be in years), while the vertical axis represents age. Lines representing the lifetimes of individuals start at their birthdate on the horizontal axis, then ascend at a 45° angle, reflecting the fact that individuals age at the rate of one year (of age) per year (of calendar time). Events during an individual's life may be represented along the lifeline — for instance, the line might

[1] These diagrams are named for Wilhelm Lexis, a 19th century statistician and demographer of many accomplishments, none of which was the invention of these diagrams, in keeping with Stigler's law of eponymy, which states that "No scientific discovery is named after its original discoverer." (cf. Christophe Vanderschrick, "The Lexis diagram, a misnomer", *Demographic Research* 4:3, pp. 97–124, http://www.demographic-research.org/Volumes/Vol4/3/.)

change colour when the individual buys an insurance policy — and the line ends at death. (Here we have marked the end with a black dot.) The collection of lifelines in a diagonal strip — individuals born at the same time (more or less broadly defined) — comprise what demographers call a "cohort". They start out together and march out along the diagonal through life, exposed to similar (or at least simultaneous) experiences. (A "cohort" was originally a unit of a Roman legion.) Note that cohorts need not be birth cohorts, as the horizontal axis of the Lexis diagram need not represent literal birthdates. For instance, a study of marriage would start "lifelines" at the date of marriage, and would refer to the "marriage cohort of 2008", for instance, while a study of student employment prospects would refer to the "student cohort of 2008", the collection of all students who completed (or started) their studies in that year.

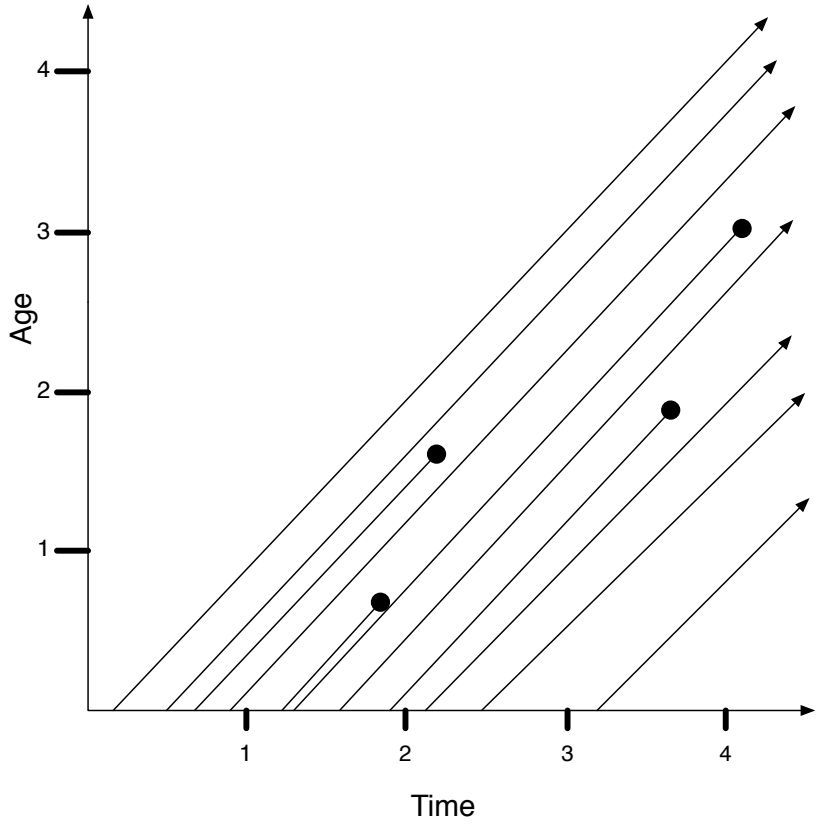


Figure 5.1: A Lexis diagram.

The census approximation involves making estimates for mortality rates in regions of the Lexis diagram. Vertical lines represent the state of the population, so a census may be represented by counting (and describing) the lifelines that cross a given vertical line. The goal is to estimate the hazard rate for a region (in age-time space) by

$$\frac{\#\text{ events}}{\text{total time at risk}}$$

The total time at risk is the total length of lifelines intersecting the region (or, to be geometric about it, the total length divided by $\sqrt{2}$), while the number of events is a count of the number of dots. The problem is that we do not know the exact total time at risk. Our censuses do tell us, though, the number of individuals at risk

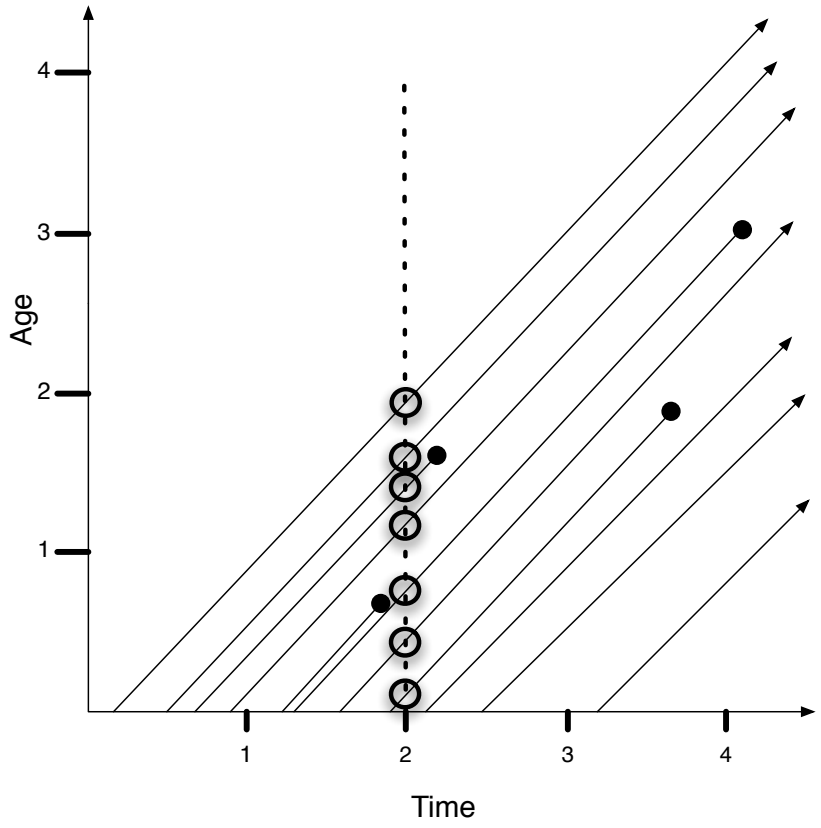


Figure 5.2: Census at time 2 represented by open circles. The population consists of 7 individuals. 4 are between ages 1 and 2, and 3 are between 0 and 1.

The count d_x described in section 5.3 tells us the number of deaths of individuals aged between x and $x+1$ (for integer x), so it is counting events in horizontal strips, such as we have shown in Figure 5.3. We are trying to estimate the central time at risk $E^c_x := \int_0^T P_{x,t}dt$, where $P_{x,t}$ is the number of individuals alive at time t whose curtate age is x. We can represent this as

$$E^c_x = \int_0^T P_{x,t}dt = \sum_{k=0}^{T-1} \mathbb{P}_{x,k}, \tag{5.7}$$

where $\mathbb{P}_{x,k}$ is defined to be the average of $P_{x,t}$ over t in the interval $[k, k+1)$. If we assume that $P_{x,t}$ is approximately linear over such an interval, we may approximate this average by $\frac{1}{2}(P(x,k) + P(x,k+1))$. Then we get the approximation

$$E^c_x = \sum_{k=0}^{T-1} \mathbb{P}_{x,k} \approx \frac{1}{2}P_{x,0} + \sum_{k=1}^{T-1} P_{x,k} + \frac{1}{2}P_{x,T}.$$

Note that this is just the trapezoid rule for approximating the integral (5.7).

Is this assumption of linearity reasonable? What does it imply? Consider first the individuals whose lifelines cross a box with lower corner (k, x). (Note that, unfortunately, the order of the age and time coordinates is reversed in the notation when we go to the geometric picture. This has no significance except sloppiness which needs to be cleaned up.) They may enter either the

left or the lower border. In the former case (corresponding to individuals born in year $x-k$) they will be counted in $P_{x,k}$; in the latter (born in $x-k+1$) case in $P_{x,k+1}$. If the births in year $x-k+1$ differ from those in year $x-k$ by a constant (that is, the difference between January 1 births in the two years is the same as the difference between February 12 births, and so on, then on average the births in the two years on a given date will contribute 1/2 year to the central years at risk, and will be counted once in the sum $P_{x,k}+P_{x,k+1}$. Important to note:

- This does not actually require that births be evenly distributed through the year.
- When we say births, we mean births that survive to age k. If those born in, say, December of one year had substantially lowered survival probability relative to a "normal" December, this would throw the calculation off.
- These assumptions are not about births and deaths in general, but rather about births and deaths of the population of interest: those who buy insurance, those who join the clinical trial, etc.

If mortality levels are low, this will suffice, since nearly all lifelines will be counted among those that cross the box. If mortality rates are high, though, we need to consider the contribution of years at risk due to those lifelines which end in the box. In this case, we do need to assume that births and deaths are evenly spread through the year. This assumption implies that conditioned on a death occurring in a box, it is uniformly distributed through the box. On the one hand, that implies that it contributes (on average) 1/4 year to the years at risk in the box. On the other hand, it implies that the probability of it having been counted in our average $\frac{1}{2}(P_{x,k}+P_{x,k+1})$ is $\frac{1}{2}$, since it is counted only if it is in the upper left triangle of box. On average, then, these should balance.

What happens when we count births and deaths only by calendar year? Note that $P'_{x,k}=P_{x,k}$ for integers k and x. One difference is that the regions in question, which are parallelograms, follow the same lifelines from the beginning of the year to the end. This makes the analysis more straightforward. Lifelines that pass through the region are counted on both ends. The other difference is that the region that begins with the census value $P_{x,k}$ ends not with $P_{x,k+1}$, but with $P_{x+1,k+1}$. Thus all the lifelines passing through the region will be counted in $P_{x,k}$ and in $P_{x+1,k+1}$, hence also in their average. This requires no further assumptions. For the lifelines that end in the region to be counted appropriately, on the other hand, requires that the deaths be evenly distributed throughout the year. (Other, slightly less restrictive assumptions, are also possible.) In this case, each death will contribute exactly 1/2 to the estimate $\frac{1}{2}(P_{x,k}+P_{x+1,k+1})$ (since it is counted only in $P_{x,k}$), and it contributes on average 1/2 year of time at risk.

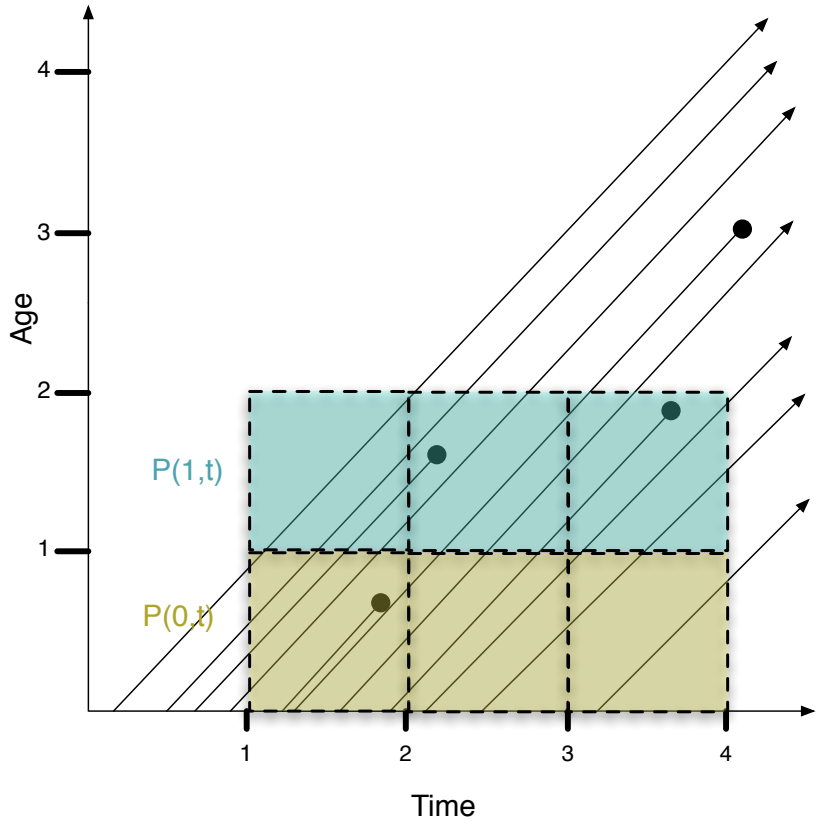


Figure 5.3: Census approximation when events are counted by actual curtate age. The vertical segments represent census counts.

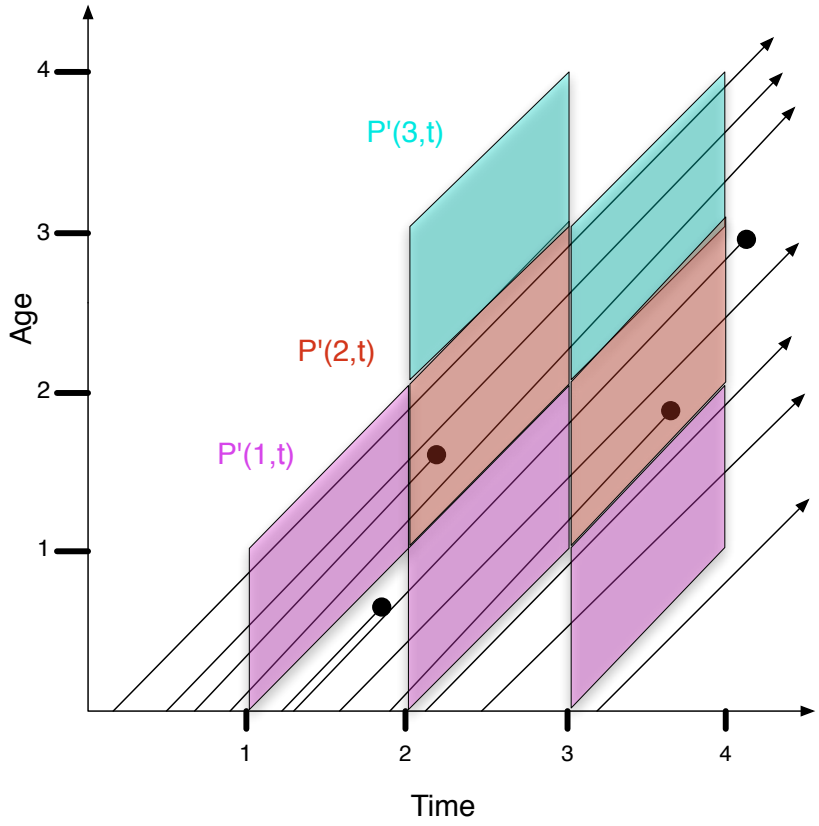


Figure 5.4: Census approximation when events are counted by calendar year of birth and death. Vertical segments bounding the coloured regions represent census counts.

Lecture 6

Comparing life tables

6.1 The binomial model

Suppose we observe n identically distributed, independent lives aged x for exactly 1 year, and record the number d_x who die. Using the notation set up for the discrete model, a life dies with probability q_x within the year.

Hence D_x, the random variable representing the numbers dying in the year conditional on n alive at the beginning of the year, has distribution

$$D_x \sim B(n, q_x)$$

giving a maximum likelihood estimator

$$\widehat{q_x} = \frac{D_x}{n}, \text{ with } \operatorname{var}(\widehat{q_x}) = \frac{q_x\,(1-q_x)}{n}$$

where using previous notation we have set $l_x = n$.

While attractively simple, this approach has significant problems. Normally failures, deaths, and other events of interest, happen continuously, even if we happen to observe or tabulate them at discrete intervals. While we get a perfectly valid estimate of q_x, the probability of an event happening in this time interval, we have no way of generalising to a question about how many individuals died in half a year, for example. And real data may be *interval truncated*: That is, the life is not under observation during the entire year, but only during the interval of ages $(x+a, x+b)$, where $0 \le a < b \le 1$. If we write D_x^i for the indicator of the event that individual i is observed to die at (curtate) age x, we have

$$P(D_x^i = 1) = {}_{b_i-a_i}q_{x+a_i}$$

Hence

$$\mathbb{E}D_x = \mathbb{E}\left(\sum_{i=1}^{n} D_x^i\right) = \sum_{i=1}^{n} {}_{b_i-a_i}q_{x+a_i}$$

There is no way to analyse (or even describe) this intra-interval refinement within the framework of the binomial model.

Nonetheless, the simplicity and tradition of the binomial model have led actuaries to develop a kind of continuous prosthetic for the binomial model, in the form of a supplemental (and hidden) model for the unobserved continuous part of the lifetime. These have been discussed in Lecture 3. In the end, these are applied through the terms Initial Exposed To Risk (E_x^0) and Central Exposed To Risk (E_x^c). These are defined more by their function than as a particular

quantity: the Initial Exposed To Risk plays the role of n in a binomial model, and E_x^c plays the role of total time at risk in an exponential model. They are linked by the **actuarial estimator**

$$E_x^0 \approx E_x^c + \frac{1}{2}d_x.$$

This may be justified from any of our fractional-lifetime models if the number of deaths is small relative to the number at risk. Thus, the actuarial estimator for q_x is

$$\widetilde{q}_x = \frac{d_x}{E_x^c + \frac{1}{2}d_x}.$$

The denominator, $E_x^c + \frac{1}{2}d_x$, comprises the observed time at risk (also called central exposed to risk) within the interval $(x, x+1)$, added to 1/2 the number of deaths (assumes deaths evenly spread over the interval). This is an estimator for E_x which is the inital exposed to risk and is what is required for the binomial model.

NB assumptions (i)-(iii) collapse to the same model, essentially (i), if $\mu_{x+\frac{1}{2}}$ is very small, since all become ${}_tq_x \approx t\mu_{x+\frac{1}{2}}$, $0 < t < 1$.

Definitions, within year $(x, x+1)$

a) E_x^c = observed total time (in years) at risk = **central exposed to risk**, with approximation $E_x^c \approx E_x - \frac{1}{2}d_x$, if required.

b) $E_x^0 (= E_x)$ = initial exposed to risk = # in risk set at age x, with approximation $E_x \approx E_x^c + \frac{1}{2}d_x$, if required.

6.2 The Poisson model

Under the assumption of a constant hazard rate (force of mortality) $\mu_{x+\frac{1}{2}}$ over the year $(x, x+1]$, we may view the estimation problem as a chain of separate hazard rate estimation problems, one for each year of life. Each individual lives some portion of a year in the age interval $(x, x+1]$, the portion being 0 (if he dies before birthday x), 1 (if he dies after birthday $x+1$), or between 0 and 1 if he dies between the two birthdays. Suppose now we lay these intervals end to end, with a mark at the end of an interval where an individual died. It is not hard to see that what results is a Poisson process on the interval $[0, E_x^c]$, where E_x^c is the total observed years at risk.

Suppose we treat E_x^c as though it were a constant. Then if D_x represents the numbers dying in the year the model uses

$$\mathrm{P}\{D_x = k\} = \frac{\left(\mu_{x+\frac{1}{2}}E_x^c\right)^k \mathrm{e}^{-\mu_{x+\frac{1}{2}}E_x^c}}{k!}, \quad k = 0, 1, 2, \cdots$$

which is an approximation to the 2-state model, and which in fact yields the same likelihood.

The estimator for the constant force of mortality over the year is

$$\widetilde{\mu}_{x+\frac{1}{2}} = \frac{D_x}{E_x^c}, \quad \text{with estimate } \frac{d_x}{E_x^c}\ .$$

Under the Poisson model we therefore have that

$$\mathrm{var}\widetilde{\mu}_{x+\frac{1}{2}} = \frac{\mu_{x+\frac{1}{2}}E_x^c}{(E_x^c)^2} = \frac{\mu_{x+\frac{1}{2}}}{E_x^c}\ .$$

So the estimate will be

$$\mathrm{var}\widetilde{\mu}_{x+\frac{1}{2}} \approx \frac{d_x}{(E_x^c)^2} \ .$$

If we compare with the **2-state stochastic model** over year $(x, x+1)$, assuming constant $\mu = \mu_{x+\frac{1}{2}}$, then the likelihood is

$$L = \prod_1^n \mu^{\delta_i} \mathrm{e}^{-\mu t_i} \ ,$$

where $\delta_i = 1$ if life i dies and $t_i = b_i - a_i$ in previous terminology (see the binomial model). Hence

$$L = \mu^{d_x} \mathrm{e}^{-\mu E_x^c}$$

and so

$$\widehat{\mu} = \frac{D_x}{E_x^c} \ .$$

The estimator is exactly the same as for the Poisson model except that both D_x and E_x^c are random variables. Using asymptotic likelihood theory we see that the estimate for the variance is

$$\mathrm{var}\widehat{\mu} \approx \frac{\mu^2}{d_x} \approx \frac{d_x}{(E_x^c)^2} \ .$$

6.3 Testing hypotheses for q_x and $\mu_{x+\frac{1}{2}}$

We note the following normal approximations:

(i) Binomial model:

$$D_x \sim B(E_x, q_x) \implies D_x \sim N(E_x q_x, E_x q_x \left(1 - q_x\right)$$

and

$$\widehat{q}_x = \frac{D_x}{E_x} \sim N\left(q_x, \frac{q_x \left(1 - q_x\right)}{E_x}\right) \ .$$

(ii) Poisson model or 2-state model:

$$D_x \sim N(E_x^c \mu_{x+\frac{1}{2}}, E_x^c \mu_{x+\frac{1}{2}})$$

and

$$\widehat{\mu}_{x+\frac{1}{2}} \sim N\left(\mu_{x+\frac{1}{2}}, \frac{\mu_{x+\frac{1}{2}}}{E_x^c}\right) \ .$$

Tests are often done using comparisons with a published **standard life table.** These can be from national tables for England and Wales published every 10 years, or insurance company data collected by the Continuous Mortality Investigation Bureau, or from other sources. (It needs to be a source

A superscript "s" denotes "from a standard table", such as q_x^s and $\mu_{x+\frac{1}{2}}^s$.

Test statistics are generally obtained from the following:

Binomial:

$$z_x = \frac{d_x - E_x q_x^s}{\sqrt{E_x q_x^s \left(1 - q_x^s\right)}} \quad \left(\approx \frac{O - E}{\sqrt{V}}\right)$$

Poisson/2-state:

$$z_x = \frac{d_x - E^c_x \mu^s_{x+\frac{1}{2}}}{\sqrt{E^c_x \mu^s_{x+\frac{1}{2}}}} \quad \left(\approx \frac{O-E}{\sqrt{V}} \right) .$$

Both of these are denoted as z_x since under a null hypothesis that the standard table is adequate $Z_x \sim N(0,1)$ approximately.

6.3.1 The tests

χ^2 test

We take

$$X = \sum_{\text{all ages } x} z_x^2$$

This gives the sum of squares of standard normal random variables under the null hypothesis and so is a sum of $\chi^2(1)$. Therefore

$$X \sim \chi^2(m) \text{ , if } m = \#\text{ years of study.}$$

H_0 : there is no difference between the standard table and the data,

H_A : they are not the same.

It is normal to use 5% significance and so the test fails if $X > \chi^2(m)_{0.95}$.

It tests large deviations from the standard table.

Disadvantages:

1. There may be a few large deviations offset by substantial agreement over part of the table. The test will not pick this up.

2. There might be bias, that is, although not necessarily large, all the deviations may be of the same sign.

3. There could be significant groups of consecutive deviations of the same sign, even if not overall.

Standardised deviations test

This tries to address point 1 above. Noting that each z_x is an observation from a standard normal distribution under H_0, the real line is divided into intervals, say 6 with dividing points at $-2, -1, 0, 1, 2$. The number of z_x in each interval is counted and compared with the expected number from a standard normal distribution. The test statistic is then

$$X = \sum_{\text{intervals}} \frac{(O-E)^2}{E} \sim \chi^2(5)$$

under the null hypothesis since this is Pearson's statistic. The problem here is that m is unlikely to be large enough to give approximate validity to the chi-square distribution. So this test is rarely appropriate.

Signs test

Test statistic X is given by

$$X = \#\{z_x > 0\}$$

Under the null hypothesis $X \sim B(m, \frac{1}{2})$, since the probability of a positive sign should be 1/2. This should be administered as a two-tailed test. It is under-powered since it ignores the size of the deviations but it will pick up small deviations of consistent sign, positive or negative, and so it addresses point 2 above.

Cumulative deviations test

This again addresses point 2 and essentially looks very similar to the logrank test between two survival curves. If instead of squaring $d_x - E_x q_x^s$ or $d_x - E_x^c \mu^s_{x+\frac{1}{2}}$, we simply sum then

$$\frac{\sum (d_x - E_x q_x^s)}{\sqrt{\sum E_x q_x^s (1 - q_x^s)}} \sim N(0,1), \quad \text{approximately}$$

and

$$\frac{\sum \left(d_x - E_x^c \mu^s_{x+\frac{1}{2}}\right)}{\sqrt{\sum E_x^c \mu^s_{x+\frac{1}{2}}}} \sim N(0,1) \quad \text{approximately.}$$

H_0 : there is no bias
H_A : there is a bias.

This test addresses point 2 again, which is that the chi-square test does not test for consistent bias.

Other tests

There are tests to deal with consecutive bias/runs of same sign. These are called the groups of signs test and the serial correlations test. Again a very large number of years, m, are required to render these tests useful.

6.3.2 An example

Table 6.3.2 presents imaginary data for men aged 90 to 95. The column ℓ_x lists the initial at risk, the number of men in the population on the census date, and d_x is the number of deaths from this initial population over the course of the year. E_x^c is the central at risk, estimated as $\ell_x - d_x/2$. Standard male British mortality for these ages is listed in column μ_x^s. (The columen $\mathring{\mu}_x$ is a graduated estimate, which will be discussed in section 6.4.

We note substantial differences between the estimates $\hat{\mu}_x$ and the standard mortality μ_x^s, but none of them is extremely large relative to the standard error: The largest z_x is 1.85. We test the two-sided alternative hypothesis, that the mortality rates in the old-people's home are different from the standard mortality rates, with a χ^2 test, adding up the z_x^2. The observed X^2 is 7.1, corresponding to an observed significance level $p = 0.31$. (Remember that we have 6 degrees of freedom, not 5, because these z_x are independent. This is not an incidence table.)

age	ℓ_x	d_x	E_x^c	$\hat{\mu}_{x+\frac{1}{2}}$	μ_x^s	z_x	$\mathring{\mu}_x$
90	40	10	35	0.29	0.202	1.1	0.25
91	35	8	31	0.258	0.215	0.52	0.28
92	22	4	18	0.20	0.236	−0.33	0.335
93	14	6	11	0.545	0.261	1.85	0.40
94	11	4	9	0.444	0.279	0.94	0.45
95	7	3	5.5	0.545	0.291	1.11	0.48

Table 6.1: Table of mortality rates for an imaginary old-people's home, with standard British male mortality given as μ_x^s, and graduated estimate $\mathring{\mu}_x$.

6.4 Graduation

Graduation is what statisticians would call "smoothing". Suppose that a company has collected its own data, producing estimates for either q_x or $\mu_{x+\frac{1}{2}}$. The estimates may be rather irregular from year to year and this could be an artefact of the population the company happens to have in a particular scheme. The underlying model should probably (but not necessarily) be smoother than the raw estimates. If it is to be considered for future predictions, then smoothing should be considered. This is called graduation.

There is always a tradeoff in smoothing procedures. Without smoothing, real patterns get lost in the random noise. Too much smoothing, though, can swamp the data in the model, so that the final estimate reflects more our choice of model than any truth gleaned from the data.

6.4.1 Parametric models

We may fit a formula to the data. Possible examples are

$$\mu_x = \mu \qquad \text{(Exponential)};$$
$$\mu_x = Be^{\theta x} \qquad \text{(Gompertz)};$$
$$\mu_x = A + Be^{\theta x} \qquad \text{(Makeham)}$$

The Gompertz can be a good model for a population of middle older age groups. The Makeham model has an extra additive constant which is sometimes used to model "intrinsic mortality", which is supposed to be independent of age. We could use more complicated formulae putting in polynomials in x.

6.4.2 Reference to a standard table

Here q_x^0, μ_x^0 represent the graduated estimates. We could have a linear dependence

$$q_x^0 = a + bq_x^s, \quad \mu_x^0 = a + b\mu_x^s$$

or possibly a translation of years

$$q_x^0 = q_{x+k}^s, \quad \mu_x^0 = \mu_{x+k}^s$$

In general there will be some assigned functional dependence of the graduated estimate on the standard table value. These are connected with the notions of accelerated lifetimes and proportional hazards, which will be central topics in the second part of the course.

6.4.3 Nonparametric smoothing

We effectively smooth our data when we impose the assumption that mortality rates are constant over a year. We may tune the strength of smoothing by requiring rates to be constant over longer intervals. This is a form of local averaging, and there are more and less sophisticated versions of this. In `Matlab` or `R` the methods available include kernel smoothing, orthogonal polynomials, cubic splines, and LOESS. These are beyond the scope of this course.

In Figure 6.1 we show a very simple example. The mortality rates are estimated by individual years or by lumping the data in five year intervals. The green line shows a moving average of the one-year estimates, in a window of width five years.

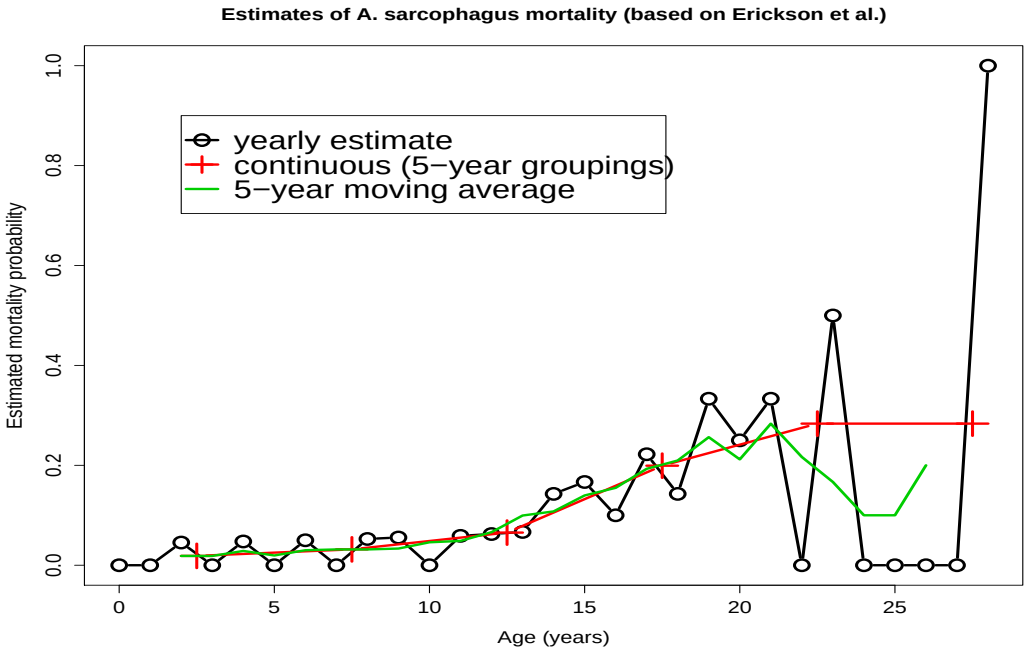


Figure 6.1: Different smooothings for *A. sarcophagus* mortality from Table 1.1.

6.4.4 Methods of fitting

1. In any of the models (binomial, Poisson, 2-state) set (say) $q_x = a + bq_x^s$ in the likelihood and use maximum likelihood estimators for the unknown parameters a, b and similarly for μ_x and other functional relationships with the standard values.

2. Use weighted least squares and minimise

$$\sum_{\text{all ages } x} w_x \left(\widehat{q}_x - q_x^0\right)^2 \quad \text{or}$$

$$\sum_{\text{all ages } x} w_x \left(\widehat{\mu}_{x+\frac{1}{2}} - \mu_{x+\frac{1}{2}}^0\right)^2$$

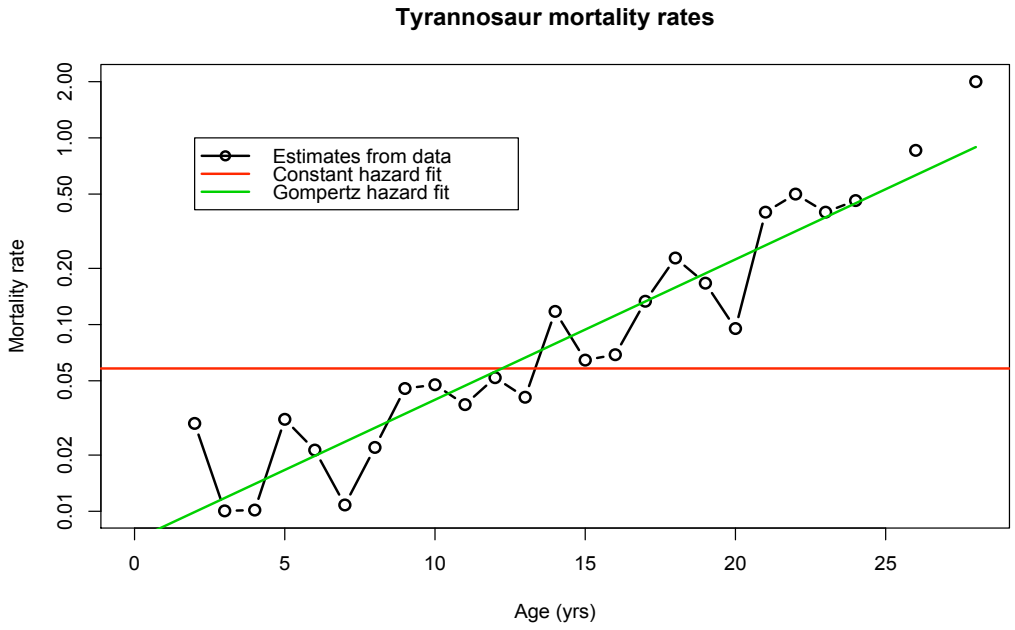


Figure 6.2: Estimated tyrannosaurus mortality rates from Table 4.3, together with exponential and Gompertz fits.

as appropriate. For the weights suitable choices are either E_x or E_x^c respectively. Alternatively we can use 1/var, where the variance is estimated for $\widehat{q}_x$ or $\widehat{\mu}_{x+\frac{1}{2}}$, respectively.

The hypothesis tests we have already covered above can be used to test the graduation fit to the data, replacing $q_x^s, \mu_{x+\frac{1}{2}}^s$ by the graduated estimates. **Note that in the χ^2 test we must reduce the degrees of freedom of the χ^2 distribution by the number of parameters estimated in the model for the graduation.** For example if $q_x^0 = a + bq_x^s$, then we reduce the degrees of freedom by 2 as the parameters a, b are estimated.

6.4.5 Examples

Standard life table

We graduate the estimates in Table 6.3.2, based on the standard mortality rates listed in the column μ_x^s, using the parametric model $\mathring{\mu}_x = a + b\mu_x^s$. The log likelihood is

$$\ell = \sum d_x \log \mathring{\mu}_{x+\frac{1}{2}} - \mathring{\mu}_{x+\frac{1}{2}} E_x^c.$$

We maximise by solving the equations

$$0 = \frac{\partial \ell}{\partial a} = \sum \left(\frac{d_x}{\hat{a} + \hat{b}\mu^s_{x+\frac{1}{2}}} - E^c_x \right)$$
$$0 = \frac{\partial \ell}{\partial b} = \sum \left(\frac{d_x \mu^s_{x+\frac{1}{2}}}{\hat{a} + \hat{b}\mu^s_x} - \mu^s_{x+\frac{1}{2}} E^c_x \right).$$

We can solve these equations numerically, to obtain $\hat{a} = -0.279$ and $\hat{b} = 2.6$. This yields the graduated estimates $\mathring{\mu}$ tabulated in the final column of Table 6.3.2. Note that these estimates have the virtue of being, on the one hand, closer to the observed data than the standard mortality rates; on the other hand smoothly and monotonically increasing.

If we had used ordinary least squares to fit the mortality rates, we would have obtained very different estimates: $\tilde{a} = -0.472$ and $\tilde{b} = 3.44$, because we would be counting the Weighted least squares, with weights proportional to E^c_x (inverse variance) solves this problem, more or less, and gives us estimates $\hat{a}^* = -0.313$ and $\hat{b}^* = 2.75$ very close to the MLE.

In Figure 6.2 we plot the mortality rate estimates for the complete population of tyrannosaurs described in Table 1.1, on a logarithmic scale, together with two parametric model fits: the exponential model, with one parameter μ estimated by

$$\hat{\mu} = \frac{1}{\bar{t}} = \frac{n}{t_1 + \cdots + t_n} \approx \frac{n}{k_1 + \cdots + k_n + n/2} = 0.058,$$

where $t_1, \ldots, t_n$ are the n lifetimes observed, and $k_i = \lfloor t_i \rfloor$ the curtate lifetimes; and the Gompertz model $\mu_s = Be^{\theta s}$, estimated by

$$\hat{\theta} \text{ solves } \frac{Q'(\hat{\theta})}{Q(\hat{\theta}) - 1} - \frac{1}{\hat{\theta}} = \bar{t},$$
$$\hat{B} := \frac{\hat{\theta}}{Q(\hat{\theta}) - 1},$$
$$\text{where } Q(\theta) := \frac{1}{n} \sum e^{\theta t_i}.$$

This yields $\hat{\theta} = 0.17$ and $\hat{B} = 0.0070$. It seems apparent to the eye that the exponential fit is quite poor, while the Gompertz fit might be pretty good. It is hard to judge the fit by eye, though, since the quality of the fit depends in part on the number of individuals at risk that go into the individual mortality-rate estimates, something which does not appear in the plot.

To test the hypothesis, we compute the predicted number of deaths in each age class $d^{(\text{exp})}_x = l_x \cdot q^{(\text{exp})}_x$ if there is a constant $\mu_x = \hat{\mu} = 0.058$, meaning that $q^{(\text{exp})}_x = 0.057$, and $d^{(\text{Gom})}_x = l_x \cdot q^{(\text{Gom})}_x$ if

$$q_x = q^{(\text{Gom})}_x := 1 - \exp\left\{ -\frac{\hat{B}}{\hat{\theta}} e^{\hat{\theta} x} \left(e^{\hat{\theta}} - 1 \right) \right\},$$

which is obtained by integrating the Gompertz hazard.

It matters little how we choose to interpret the deviations in the column $z^{(\text{exp})}_x$ — with values going up as high as 6.65, it is clear that these could not have come from a normal distribution, and we must reject the null hypothesis that these lifetimes came from an exponential distribution.

age	l_x	d_x	$q_x^{(\text{exp})}$	$d_x^{(\text{exp})}$	$z_x^{(\text{exp})}$	$q_x^{(\text{Gom})}$	$d_x^{(\text{Gom})}$	$z_x^{(\text{Gom})}$
0	103	0	0.057	5.8	-2.48	0.007	0.7	-0.97
1	103	0	0.057	5.8	-2.48	0.009	0.9	-0.97
2	103	3	0.057	5.8	-1.20	0.011	1.1	1.81
3	100	1	0.057	5.7	-2.02	0.013	1.3	-0.25
4	99	1	0.057	5.6	-2.00	0.015	1.5	-0.41
5	98	3	0.057	5.5	-1.11	0.018	1.8	0.94
6	95	2	0.057	5.4	-1.50	0.021	2.0	-0.02
7	93	1	0.057	5.3	-1.91	0.025	2.4	-0.90
8	92	2	0.057	5.2	-1.45	0.030	2.8	-0.47
9	90	4	0.057	5.1	-0.50	0.036	3.2	0.45
10	86	4	0.057	4.9	-0.40	0.042	3.6	0.19
11	82	3	0.057	4.6	-0.78	0.050	4.1	-0.56
12	79	4	0.057	4.5	-0.23	0.059	4.7	-0.32
13	75	3	0.057	4.2	-0.62	0.070	5.3	-1.02
14	72	8	0.057	4.1	2.00	0.083	6.0	0.87
15	64	4	0.057	3.6	0.21	0.098	6.2	-0.95
16	60	4	0.057	3.4	0.34	0.115	6.9	-1.17
17	56	7	0.057	3.2	2.22	0.135	7.6	-0.22
18	49	10	0.057	2.8	4.47	0.159	7.8	0.87
19	39	6	0.057	2.2	2.63	0.186	7.2	-0.51
20	33	3	0.057	1.9	0.85	0.216	7.1	-1.75
21	30	10	0.057	1.7	6.56	0.252	7.6	1.03
22	20	8	0.057	1.1	6.65	0.292	5.8	1.07
23	12	4	0.057	0.7	4.15	0.336	4.0	-0.02
24	8	3	0.057	0.5	3.90	0.386	3.1	-0.06
25	5	0	0.057	0.3	-0.55	0.440	2.2	-1.98
26	5	3	0.057	0.3	5.26	0.498	2.5	0.46
27	2	0	0.057	0.1	-0.35	0.559	1.1	-1.59
28	2	2	0.057	0.1	5.78	0.623	1.2	1.10

Table 6.2: Life table for tyrannosaurs, with fit to exponential and Gompertz models, and based on data from Table 1.1.

As for the Gompertz model, the deviations are all quite moderate. We compute $\sum z_x^2 = 26.1$. There are 29 categories, but we have estimated 2 parameters, so this needs to be compared to the χ^2 distribution with 27 degrees of freedom. The cutoff for a test at the 0.05 level is 40.1, so we do not reject the null hypothesis.

As you have already learned, the χ^2 approximation doesn't work very well when the expected numbers in some categories are too low. This is certainly the case in this case, with d_x^{Gom} as low as 0.7. (That is, we are using a normal approximation with mean 0.7 for a quantity which takes on integer values. That obviously cannot be right.) The solution is to lump categories together. If we replace the first 10 years by a single category, it will have an expected number of deaths equal to $0.171 \cdot 103 = 17.7$, as compared with exactly 17 observed deaths, producing a z value of 0.18. Similarly, we cut off the final three years (since collectively they correspond to the certain event that all remaining individuals die), leaving us with $\sum z_x^2 = 11$ on 14 degrees of freedom. Again, this is a perfectly ordinary value of the χ^2 variable, and we do not reject the hypothesis

of Gompertz mortality.

Lecture 7

Multiple decrements model

Reading: Gerber Sections 7.1-7.3 and 11.7, Cox-Oakes Sections 9.1-9.2, Norris Section 1.10, CT4 Units 9,10-3,10-4
Further reading: Cox-Oakes Section 9.3

7.1 The Poisson model

Under the assumption of a constant hazard rate (force of mortality) $\mu_{x+\frac{1}{2}}$ over the year $(x, x+1]$, we may view the estimation problem as a chain of separate hazard rate estimation problems, one for each year of life. Each individual lives some portion of a year in the age interval $(x, x+1]$, the portion being 0 (if he dies before birthday x), 1 (if he dies after birthday $x+1$), or between 0 and 1 if he dies between the two birthdays. Suppose now we lay these intervals end to end, with a mark at the end of an interval where an individual died. It is not hard to see that what results is a Poisson process on the interval $[0, E^c_x]$, where E^c_x is the total observed years at risk.

Suppose we treat E^c_x as though it were a constant. Then if D_x represents the numbers dying in the year the model uses

$$\mathrm{P}\big\{D_x = k\big\} = \frac{\left(\mu_{x+\frac{1}{2}} E^c_x\right)^k \mathrm{e}^{-\mu_{x+\frac{1}{2}} E^c_x}}{k!}, \quad k = 0, 1, 2, \cdots$$

which is an approximation to the 2-state model, and which in fact yields the same likelihood.

The estimator for the constant force of mortality over the year is

$$\widetilde{\mu}_{x+\frac{1}{2}} = \frac{D_x}{E^c_x}, \quad \text{with estimate } \frac{d_x}{E^c_x} \ .$$

Under the Poisson model we therefore have that

$$\mathrm{var}\widetilde{\mu}_{x+\frac{1}{2}} = \frac{\mu_{x+\frac{1}{2}} E^c_x}{(E^c_x)^2} = \frac{\mu_{x+\frac{1}{2}}}{E^c_x} \ .$$

So the estimate will be

$$\mathrm{var}\widetilde{\mu}_{x+\frac{1}{2}} \approx \frac{d_x}{(E^c_x)^2} \ .$$

If we compare with the **2-state stochastic model** over year $(x, x+1)$, assuming constant $\mu = \mu_{x+\frac{1}{2}}$, then the likelihood is

$$L = \prod_1^n \mu^{\delta_i} \mathrm{e}^{-\mu t_i} \ ,$$

where $\delta_i = 1$ if life i dies and $t_i = b_i - a_i$ in previous terminology (see the binomial model). Hence

$$L = \mu^{d_x} \mathrm{e}^{-\mu E_x^c}$$

and so

$$\widehat{\mu} = \frac{D_x}{E_x^c} \ .$$

The estimator is exactly the same as for the Poisson model except that both D_x and E_x^c are random variables. Using asymptotic likelihood theory we see that the estimate for the variance is

$$\mathrm{var}\widehat{\mu} \approx \frac{\mu^2}{d_x} \approx \frac{d_x}{(E_x^c)^2} \ .$$

Are we justified in treating E_x^c as though it were fixed? Certainly it's not exactly the same: The numerator and denominator are both random, and they are not even independent. One way of looking at this is to ask, how different would our estimate have been in a given realisation, had we fixed the total time under observation in advance. If we observe m lives from the start of year x, we see that D_x is approximately normal with mean mq_x and variance $mq_x(1-q_x)$, while E_x^c is normal with mean $m - mq_x(1-e_*)$, where e_* is the expected remaining length of a life starting from age x, conditioned on its being less than 1; and variance $m\sigma^2$, where σ^2 is the variance in time under observation of a single life. (If μ_x is not very large, e_* is close to $\frac{1}{2}$.) Looking at the first-order Taylor series expansion, we see that the ratio D_x/E_x^c varies only by a normal error term times $m^{-1/2}$, plus a bias of order m^{-1}. For large m, then, the estimate on the basis of fixed m (number of individuals) is almost the same as the estimate we would have made from observing the Poisson model for the fixed total time at risk $m - mq_x(1-e_*)$.

7.2 Rates in the single decrement model

The rate parameter in the two-state Markov model with $L \sim \mathrm{Exp}(\lambda)$ has the infinitesimal interpretation

$$\mathbb{P}(X_{t+\varepsilon} = 1 | X_t = 0) = \mathbb{P}(L \le t+\varepsilon | L > t) = 1 - e^{-\lambda\varepsilon} = \lambda\varepsilon + o(\varepsilon), \tag{7.1}$$

and for a general L with right-continuous density and hence right-continuous force of mortality $t \mapsto \mu_t$, we have

$$\mathbb{P}(X_{t+\varepsilon} = 1 | X_t = 0) = \mathbb{P}(L \le t+\varepsilon | L > t) = 1 - \exp\left\{-\int_t^{t+\varepsilon} \mu_s ds\right\} = \mu_t\varepsilon + o(\varepsilon), \tag{7.2}$$

since, by l'Hôpital's rule

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\left(1 - \exp\left\{-\int_t^{t+\varepsilon} \mu_s ds\right\}\right) = \lim_{\varepsilon\downarrow 0} \mu_{t+\varepsilon} \exp\left\{-\int_t^{t+\varepsilon} \mu_s ds\right\} = \mu_t. \tag{7.3}$$

It is therefore natural to express the two-state model by a time-dependent Q-matrix

$$Q(t) = \begin{pmatrix} -\lambda(t) & \lambda(t) \\ 0 & 0 \end{pmatrix}, \qquad \text{where } \lambda(t) = \mu_t = h_L(t). \tag{7.4}$$

For estimation purposes, it has been convenient to add as additional assumption that $\lambda(t) = \mu_t = \mu_{x+\frac{1}{2}} = \lambda(x+\frac{1}{2})$ is constant on $x \le t < x+1$, $x \in \mathbb{N}$.

We have expressed the process $X = (X_t)_{t\geq 0}$ as $X_t = 0$ for $0 \leq t < L$ and $X_t = 1$ for $t > L$, where L is the transition time. Given the observed transition times $y_1, \ldots, y_n$ of n independent copies of X (corresponding to n different 'individuals'), we have constructed two different sets of maximum likelihood estimates

$$\hat{\mu}^{(0)}_{x+\frac{1}{2}}(y_1, \ldots, y_n) = -\ln\left(\hat{q}^{(0)}_x(y_1, \ldots, y_n)\right) = \ln\left(1 - \frac{d_x(y_1, \ldots, y_n)}{\ell_x(y_1, \ldots, y_n)}\right),$$
$$\hat{\mu}_{x+\frac{1}{2}}(y_1, \ldots, y_n) = \frac{d_x(y_1, \ldots, y_n)}{\tilde{\ell}_x(y_1, \ldots, y_n)}, \qquad 0 \leq x \leq \max\{y_1, \ldots, y_n\}.$$

If we furthermore assume that $\lambda(t) \equiv \lambda$ for all $t \geq 0$, then the maximum likelihood estimator is simply

$$\hat{\lambda} = \frac{n}{y_1 + \ldots + y_n} = \frac{d_0 + \ldots + d_{[\max\{y_1,\ldots,y_n\}]}}{\tilde{\ell}_0 + \ldots + \tilde{\ell}_{[\max\{y_1,\ldots,y_n\}]}}. \tag{7.5}$$

7.3 Multiple decrement models

The simplest (and most immediately fruitful) way to generalise the single-decrements model is to allow transitions to multiple absorbing states. Of course, as demographer Kenneth Wachter has put it, it may seem peculiar to introduce multiple "dead" states into our models since there is only one way of being dead; but (as he continues), there are many ways of getting there. Further, there are many other settings which can be modelled by a single nonabsorbing state transitioning into one of several possible absorbing states. Some examples are

- A working population insured for disability might transition into multiple different possible causes of disability, which may be associated with different costs.
- Workers may leave a company through retirement, resignation, or death.
- A model of unmarried cohabitations, which may end either by separation or marriage.
- Unemployed individuals may leave that state either by finding a job, or by giving up looking for work and so becoming "long-term unemployed".

An important common element is that calling the states "absorbing" does not have to mean that it is a deathlike state, from which nothing more happens. Rather, it simply means that our model does not follow any further developments.

7.3.1 An introductory example

This example is taken from section 8.2 of [Wac].

According to United Nations statistics, the probability of dying for men in Zimbabwe in 2000 was ${}_5q_{30} = 0.1134$, with AIDS accounting for approximately 4/5 of the deaths in this age group. Suppose we wish to answer the question: what would be the effect on mortality rates of a complete cure for AIDS?

One might immediately be inclined to think that the mortality rate would be reduced to 1/5 of its current rate, so that what the probability of dying of some other cause in the absence of AIDS, which we might write as ${}_5q^{OTHER*}_{30}$, would be 0.02268. On further reflection, though, it seems that this is too low: This is the proportion of people aged 30 who *currently* die of causes

other than AIDS. If AIDS were eliminated, surely some of the people who now die of AIDS would instead die of something else.

Of course, this is not yet a well-defined mathematical problem. To make it such, we need to impose extra conditions. In particular, we impose the **competing risks** assumption: Individual causes of death are assumed to act independently. You might imagine an individual drawing lots from multiple urns, labelled "AIDS", "Stroke", "Plane crash", to determine whether he will die of this cause in the next year. The fraction of black lots among the white is precisely q_x, when the individual has age x. If he gets no black lot, he survives the year. If he draws two or more, we only get to see the one drawn first, since he can only die once. The probability of surviving is then the product of the survival probabilities:

$$ {}_iq_x = 1 - \left(1 - {}_iq_x^{CAUSE1}\right)\left(1 - {}_iq_x^{CAUSE2}\right)\cdots \tag{7.6} $$

What is the fraction of deaths due to a given cause? Assuming constant mortality rate over the time interval due to each cause, we have

$$ 1 - {}_tq_x^{CAUSE1} = e^{-t\lambda_x^{CAUSE1}}. $$

Given a death, the probability of it being due to a given cause, is proportional to the associated hazard rate. Consequently,

$$ \lambda_x^{CAUSE1} = \text{fraction of deaths due to CAUSE 1} \times \lambda_x, $$

which implies that

$$ {}_tq_x^{CAUSE1} = 1 - (1 - {}_tq_x)^{\text{fraction of deaths due to CAUSE 1}}. $$

(Note that this is the same formula that we use for changing lengths of time intervals: ${}_tq_x = 1 - (1 - {}_1q_x)^t$.) This tells us the probability of dying from cause 1 in the absence of any other cause. The probability of dying of any cause at all is then given by (7.6).

Applying this to our Zimbabwe AIDS example, treating the causes as being either AIDS or OTHER, we see that the probability of dying of AIDS in the absence of any other cause is

$$ {}_5q_{30}^{AIDS*} = 1 - (1 - {}_5q_{30})^{4/5} = 1 - 0.8866^{4/5} = 0.0918, $$

while the probability of dying of any other cause, in the absence of AIDS, is

$$ {}_5q_{30}^{OTHER*} = 1 - (1 - {}_5q_{30})^{4/5} = 1 - 0.8866^{1/5} = 0.0238. $$

Appropriately, we have the total cause of death $0.1138 = 1 - (1 - 0.0918)(1 - 0.0238)$.

Is the competing risks assumption reasonable? Another way of putting this is to ask, what circumstances would cause the assumption to be violated? The answer is: Competing risks is violated when a subpopulation is at higher than average risk for multiple causes of death simultaneously; or conversely, when those at higher than average risk for one cause of death are protected from another cause of death. For example, smokers have more than 10 times the risk of dying from lung cancer that nonsmokers have; but they also have substantially higher mortality from other cancers, heart disease, stroke, and so on. If a perfect cure for lung cancer were to be found, it would not save nearly as many lives as one might suppose, from a competing-risks calculation like the one above, because the lives that would be saved would be almost all those of smokers, and they would be more likely to die of something else than an equivalent number of saved lives from the general population.

7.3.2 Basic theory

We consider here more general $1+m$-state Markov models with state space $\mathbb{S} = \{0, \ldots, m\}$ that only have one transition from 0 to j, for some $1 \leq j \leq m$, with absorption in j. We can write down an – in general time-dependet – Q-matrix

$$Q(t) = \begin{pmatrix} -\lambda_+(t) & \lambda_1(t) & \cdots & \lambda_m(t) \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \qquad \text{where } \lambda_+(t) = \lambda_1(t) + \ldots + \lambda_m(t). \tag{7.7}$$

Such models occur naturally where insurance policies provide different benefits for different causes of death, or distinguish death and disability, possibly in various different strengths or forms. This is also clearly a building block (one transition only) for general Markov models, where states $j = 1, \ldots, m$ may not all be absorbing.

Such a model depends upon the assumption that different causes of death act independently — that is, the probability of dying is the product of what might be understood as the probability of dying from each individual cause acting alone.

7.3.3 Multiple decrements – time-homogeneous rates

In the time-homogenous case, we can think of the multiple decrement model as m exponential clocks C_j with parameters λ_j, $1 \leq j \leq m$, and when the first clock goes off, say, clock j, the only transition takes place, and leads to state j. Alternatively, we can describe the model as consisting of one $L \sim \text{Exp}(\lambda_+)$ holding time in state 0, after which the new state j is chosen *independently* with probability λ_j/λ_+, $1 \leq j \leq m$. The likelihood for a sample of size 1 consists of two ingredients, the density $\lambda_+ e^{-t\lambda_+}$ of the exponential time, and the probability λ_j/λ_+ of the transition observed. This gives $\lambda_j e^{-t\lambda_+}$, or, for a sample of size n of lifetimes t_i and states j_i, $1 \leq i \leq n$,

$$\prod_{i=1}^{n} \lambda_{j_i} e^{-t_i\lambda_+} = \prod_{j=1}^{m} \lambda_j^{n_j} e^{-\lambda_j(t_1+\ldots+t_n)}, \tag{7.8}$$

where n_j is the number of transitions to j. Again, this can be solved factor by factor to give

$$\hat{\lambda}_j = \frac{n_j}{t_1 + \ldots + t_n}, \qquad 1 \leq j \leq m. \tag{7.9}$$

In particular, we find again $\hat{\lambda}_+ = n/(t_1 + \ldots + t_n)$, since $n_1 + \ldots + n_m = n$.

In the competing-clocks description, we can interpret the likelihood as consisting of m ingredients, namely the density $\lambda_j e^{-\lambda_j t}$ of clock j to go off at time t, and probabilities $e^{-\lambda_k t}$ of clocks C_k, $k \neq j$, to go off after time t.

Lecture 8

Multiple Decrements: Theory and Examples

8.1 Estimation for general multiple decrements

We can deduce from either description in the previous section that the likelihood for a sample of n independent lifetimes $y_1, \ldots, y_n$ and respective new states $j_1, \ldots, j_n$, each (y_i, j_i) sampled from (L, J), is given by

$$\prod_{i=1}^{n} \lambda_{j_i}(y_i) \exp\left\{ -\int_0^{y_i} \lambda_+(t)dt \right\}. \tag{8.1}$$

Let us assume that the forces of decrement $\lambda_j(t) = \lambda_j(x + \frac{1}{2})$ are constant on $x \le t < x + 1$, for all $x \in \mathbb{N}$ and $1 \le j \le m$. Then the likelihood can be given as

$$\prod_{x \in \mathbb{N}} \prod_{j=1}^{m} \left(\lambda_j(x + \tfrac{1}{2})\right)^{d_{j,x}} \exp\left\{ -\tilde{\ell}_x \lambda_j(x + \tfrac{1}{2}) \right\}, \tag{8.2}$$

where $d_{j,x}$ is the number of decrements to state j between ages x and $x + 1$, and $\tilde{\ell}_x$ is the total time spent alive between ages x and $x + 1$.

Now the parameters are $\lambda_j(x + \frac{1}{2})$, $x \in \mathbb{N}$, $1 \le j \le m$, and they are again well separated to deduce

$$\hat{\lambda}_j(x + \tfrac{1}{2}) = \frac{d_{j,x}}{\tilde{\ell}_x}, \qquad 1 \le j \le m, \quad 0 \le x \le \max\{L_1, \ldots, L_n\}. \tag{8.3}$$

Similarly, we can try to adapt the method to get maximum likelihood estimators from the curtate lifetimes. We can write down the likelihood as

$$\prod_{i=1}^{n} p_{(J,K)}(j_i, [y_i]) = \prod_{x \in \mathbb{N}} (1 - q_x)^{\ell_x - d_x} \prod_{j=1}^{m} q_{j,x}^{d_{j,x}}, \tag{8.4}$$

but $1 - q_x = 1 - q_{1,x} - \ldots - q_{m,x}$ does not factorise, so we have to maximise simultaneously for all $1 \le j \le m$ expressions of the form

$$(1 - q_1 - \ldots - q_m)^{\ell - d_1 - \ldots - d_m} \prod_{j=1}^{m} q_j^{d_j}. \tag{8.5}$$

(We suppress the indices x.) A zero derivative with respect to q_j amounts to

$$(\ell - d_1 - \ldots - d_m) q_j = d_j (1 - q_1 - \ldots - q_m), \qquad 1 \le j \le m, \tag{8.6}$$

and summing over j gives

$$(\ell - d)q = d(1-q) \qquad \Rightarrow \qquad q = \frac{d}{\ell}. \tag{8.7}$$

and then

$$(\ell - d)q_j = d_j(1-q) \qquad \Rightarrow \qquad q_j = \frac{d_j(1-q)}{\ell - d} = \frac{d_j}{\ell} \tag{8.8}$$

so that, if we display the suppressed indices x again,

$$\hat{q}^{(0)}_{j,x} = \hat{q}^{(0)}_{j,x}(y_1, j_1, \ldots, y_n, j_n) = \frac{d_{j,x}}{\ell_x}. \tag{8.9}$$

Now we've done essentially all maximum likelihood calculations. This one was the only one that was not totally trivial. At repeated occurrences of the same factors, we have been and will be less explicit about these calculations. We'll derive likelihood functions, note that they factorise and identify the factors as being of one of the three forms

$$\begin{aligned} (1-q)^{\ell-d}q^d \qquad &\Rightarrow \qquad \hat{q} = d/\ell \\ \mu^d e^{-\mu\ell} \qquad &\Rightarrow \qquad \hat{\mu} = d/\ell \\ (1 - q_1 - \ldots - q_m)^{\ell - d_1 - \ldots - d_m} \prod_{j=1}^{m} q_j^{d_j} \qquad &\Rightarrow \qquad \hat{q}_j = d_j/\ell, \quad j = 1, \ldots, m. \end{aligned}$$

and deduce the estimates.

8.2 Example: Workforce model

A company is modelling its workforce using the model

$$Q(t) = \begin{pmatrix} -\lambda(t) - \sigma(t) - \mu(t) & \lambda(t) & \sigma(t) & \mu(t) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{8.10}$$

with four states $\mathbb{S} = \{W, V, I, \Delta\}$, where W ='working', V ='left the company voluntarily', I ='left the company involuntarily' and Δ ='left the company through death'.

If we observe n_x people aged x, then

$$\hat{\lambda}_{x+\frac{1}{2}} = \frac{d_{x,V}}{\tilde{\ell}_x}, \quad \hat{\sigma}_{x+\frac{1}{2}} = \frac{d_{x,I}}{\tilde{\ell}_x}, \quad \hat{\mu}_{x+\frac{1}{2}} = \frac{d_{x,\Delta}}{\tilde{\ell}_x} \tag{8.11}$$

where $\tilde{\ell}_x$ is the total amount of time spent working aged x, $d_{x,V}$ is the total number of workers who left the company voluntarily aged x, $d_{x,I}$ is the total number of workers who left the company involuntarily aged x, $d_{x,\Delta}$ is the total number of workers dying aged x.

Lecture 9

Multiple decrements: The distribution of the endpoint

9.1 Which state do we end up in?

The time-homogeneous multiple-decrement model makes a transition at the minimum of m exponential clocks as opposed to one clock in the single decrement model. In the same way, we can construct the time-inhomogeneous multiple-decrement model from m independent clocks C_j with hazard function $\lambda_j(t)$, $1 \le j \le m$. Then the likelihood for a transition at time t to state j is the product of $f_{C_j}(t)$ and $\bar{F}_{C_k}(t)$.

By Exercise A.1.2, the hazard function of $L = \min\{C_1, \ldots, C_m\}$ is given by $h_L(t) = h_{C_1}(t) + \ldots + h_{C_m}(t) = \lambda_1(t) + \ldots + \lambda_m(t) = \lambda_+(t)$, and we can also calculate

$$\begin{aligned}
\mathbb{P}(L = C_j | L = t) &\lim_{\varepsilon \downarrow 0} \frac{\mathbb{P}(t \le C_j < t+\varepsilon, \min\{C_k : k \ne j\} \ge C_j)}{\mathbb{P}(t \le L < t+\varepsilon)} \\
&\begin{cases} \ge & \lim\limits_{\varepsilon \downarrow 0} \dfrac{\frac{1}{\varepsilon}\mathbb{P}(t \le C_j < t+\varepsilon, \min\{C_k : k \ne j\} \ge t+\varepsilon)}{\frac{1}{\varepsilon}\mathbb{P}(t \le L < t+\varepsilon)} \\ \le & \lim\limits_{\varepsilon \downarrow 0} \dfrac{\frac{1}{\varepsilon}\mathbb{P}(t \le C_j < t+\varepsilon, \min\{C_k : k \ne j\} \ge t)}{\frac{1}{\varepsilon}\mathbb{P}(t \le L < t+\varepsilon)} \end{cases} \\
&= \frac{f_{C_j}(t) \prod_{k \ne j} \bar{F}_{C_k}(t)}{f_L(t)} = \frac{h_{C_j}(t)}{h_L(t)} = \frac{\lambda_j(t)}{\lambda_+(t)},
\end{aligned}$$

and we obtain

$$\mathbb{P}(L = C_j) = \int_0^\infty \mathbb{P}(L = C_j | L = t) f_L(t) dt = \int_0^\infty \lambda_j(t) \bar{F}_L(t) dt = \mathbb{E}(\Lambda_j(L)), \tag{9.1}$$

where $\Lambda_j(t) = \int_0^t \lambda_j(s) ds$ is the integrated hazard function. (For the last step we used that $\mathbb{E}(g(L)) = \int_0^\infty g'(t) \bar{F}_L(t) dt$ for all increasing differentiable $g : [0, \infty) \to [0, \infty)$ with $g(0) = 0$.)

The discrete (curtate) lifetime model: We can also split the curtate lifetime $K = [L]$ according to the type of decrement J ($J = j$ if $L = T_j$) and define

$$q_{j,x} = \mathbb{P}(L < x+1, J = j | L > x), \qquad 1 \le j \le m, \quad x \in \mathbb{N}, \tag{9.2}$$

then clearly for $x \in \mathbb{N}$

$$q_{1,x} + \ldots + q_{m,x} = q_x \tag{9.3}$$

and, for $1 \leq j \leq m$,

$$p_{(J,K)}(j,x) = \mathbb{P}(J = j, K = x) = \mathbb{P}(L \leq x+1, J = j | L > x)\mathbb{P}(L > x) = p_0 \dots p_{x-1} q_{j,x}. \quad (9.4)$$

Note that this bivariate probability mass function is simple, whereas the joint distribution of (L, J) is conceptually more demanding since L is continuous and J is discrete. We chose to express the marginal probability density function of L and the conditional probability mass function of J given $L = t$. In the assignment questions, you will see an alternative description in terms of sub-probability densities $g_j(t) = \frac{d}{dt}\mathbb{P}(L \leq t, J = j)$, which you can normalise – $g_j(t)/\mathbb{P}(J = j)$ is the conditional density of L given $J = j$.

9.2 Cohabitation dissolution model

There has been considerable interest in the influence of nonmarital birth on the likelihood of a child growing up without one of its parents. In the paper [Kie01] relevant data are given for nine different western European countries. We give a summary of some of the UK data in Table 9.1. We represent the data in terms of a multiple decrement model in which the one nonabsorbing state is cohabitation, and this leads to the two absorbing states, which are marriage or separation. (Of course, there is a third absorbing state, corresponding to the death of one of the partners, but this did not appear in the data. And of course, the marriage state is not actually absorbing, except in fairy tales. A more complete analysis would splice this model onto a model of the fate of marriages. There are data in the article on rates of separation after marriage, for those interested in filling out the model.) Time, in this model, begins with the birth of the first child. Because of the way the data are given, we treat the hazard rates as constant in the time intervals $[0, 1]$, $[1, 3]$, and $[3, 5]$. There are no data about what happens after 5 years. We write d_x^M and d_x^S for the number of individuals marrying and separating, respectively, and similar for the estimation of hazard rates. (For simplicity, we have divided the separation data, which were actually only given for the periods $[0, 3]$ and $[3, 5]$, as though there were a count for separations in $[0, 1]$.)

Table 9.1: Data from [Kie01] on rates of conversion of cohabitations into marriage or separation, by years since birth of first child

n	after 3 years	after 5 years
106	61	48

(a) % cohabiting couples remaining together (from among those who did not marry)

n	1 year	3 years	5 years
150	18	30	39

(b) % of cohabiting couples who marry within stated time.

Translating the data in Table 9.1 into a multiple-decrement life table requires some interpretive work.

1. There are only 106 individuals given for the data on separation; this is both because the individuals who eventually married were excluded from this tabulation, and because the two tabulations were based on slightly different samples.

2. The data are given in percentages.

3. There is no count of separations in the first year.

4. Note that separations are given by survival percentages, while marriages are given by loss percentages.

We now construct a combined life table from the data in Table 9.1. The purpose of this model is to integrate information from the two data sets. This requires some assumptions, to wit, that the transition rates to the two different absorbing states are the same for everyone, and that they are constant over the periods 0–1, 1–3, 3–5 (and constant over 0–3 for separation).

The procedure is essentially the same as the construction of the single-decrement life table, except that the survival is decremented by both loss counts d_x^M and d_x^S; and the estimation of years at risk $\tilde{\ell}_x$ now depends on both decrements, so is

$$\tilde{\ell}_x = \ell_{x'}(x' - x) + (d_x^M + d_x^S)\frac{x' - x}{2},$$

where x' is the next age on the life table. Thus, for example, $\tilde{\ell}_1$, which is the number of years at risk from age 1 to 3, is $64 \cdot 2 + 41 \cdot 1 = 169$.

One of the challenges is that we observe transitions to Separation conditional on never being Married, but the transitions to Married are unconditioned. The data for marriage are more straightforward: These are absolute decrements, rather than conditional ones. If we set up a life table an a radix of 1000, we know that the decrements due to marriage should be exactly the percentages given in Table 9.2(b); that is, 180, 120, and 90. We begin by putting these into our multiple-decrements life table, Table 9.2.

Of these nominal 1000 individuals, there are 610 who did not marry. Multiplying by the percentages in Table 9.2(a) we estimate 238 separations in the first 3 years, and 79 in the next 2 years.

Table 9.2: Multiple decrement life table for survival of cohabiting relationships, from time of birth of first child, computed from data in Table 9.1.

x	ℓ_x	d_x^M	d_x^S	$\tilde{\ell}_x$	$\hat{\mu}_x^M$	$\hat{\mu}_x^S$
0–1	1000	180	?	?	?	?
1–3	?	120	?	?	?	?
3–5	462	90	79	755	?	?

At this point, the only thing preventing us from completing the life table is that we don't know how to allocate the 238 separations between the first two rows, which makes it impossible to compute the total years at risk for each of these intervals. The easiest way is to begin by making a first approximation by assuming the marriage rate is the unchanged over the first three years, and then coming back to correct it. Our first step, then, is the multiple-decrement table in Table 9.3. According to our usual approximation, we estimate the years at risk for the first period as $1000 \cdot 3 - 538 \cdot 1.5 = 2193$; and in the second period as $462 \cdot 2 - 169 \cdot 1 = 755$. The two decrement rates are then estimated by dividing the number of events by the number of years at risk.

Table 9.3: First approximation multiple decrement table.

x	ℓ_x	d_x^M	d_x^S	$\tilde{\ell}_x$	$\hat{\mu}_x^M$	$\hat{\mu}_x^S$
0–3	1000	300	238	2193	0.137	0.109
3–5	462	90	79	755	0.119	0.105

What correction do we need? We need to estimate the number of separations in the first year. Our model says that we have two independent times S and M, the former with constant hazard rate $\mu_{1.5}^S$ on $[0,3]$, the other with rate $\mu_{\frac{1}{2}}^M$ on $[0,1]$ and μ_2^M on $[1,3]$. Of the 238 separations in $[0,3]$, we expect the fraction in $[0,1]$ to be about

$$\mathbb{P}\big\{S<1\,\big|\,S<\min\{3,M\}\big\}=\frac{\mathbb{P}\{\min\{S,M\}<1\,\&\,S<M\}}{\mathbb{P}\{\min\{S,M\}<1\,\&\,S<M\}+\mathbb{P}\{1\le\min\{S,M\}<3\,\&\,S<M\}}.$$

Since $\min\{S,M\}$ is exponential with rate $\mu_{\frac{1}{2}}^M+\mu_{1.5}^S$ on $[0,1]$ and rate $\mu_2^M+\mu_{1.5}^S$ on $[1,3]$, and independent of $\{S<M\}$, we have

$$\mathbb{P}\{\min\{S,M\}<1\,\&\,S<M\}=\frac{\mu_{1.5}^S}{\mu_{\frac{1}{2}}^M+\mu_{1.5}^S}\left(1-e^{-\mu_{\frac{1}{2}}^M-\mu_{1.5}^S}\right);$$

$$\mathbb{P}\{1<\min\{S,M\}<3\,\&\,S<M\}=\frac{\mu_{1.5}^S}{\mu_2^M+\mu_{1.5}^S}e^{-\mu_{\frac{1}{2}}^M-\mu_{1.5}^S}\left(1-e^{-2\mu_2^M-2\mu_{1.5}^S}\right).$$

In Table 9.3 we have taken both values μ^M to be equal, the estimate being 0.137; were we to do a second round of correction we could take the estimates to be distinct. We thus obtain

$$\mathbb{P}\{\min\{S,M\}<1\,\&\,S<M\}\approx\frac{0.109}{0.137+0.109}\left(1-e^{-0.137-0.109}\right)=0.0966;$$

$$\mathbb{P}\{1<\min\{S,M\}<3\,\&\,S<M\}\approx\frac{0.109}{0.137+0.109}e^{-0.137-0.109}\left(1-e^{-2\cdot0.137-2\cdot0.109}\right)=0.135.$$

Thus, we allocate $0.418*238=99.4$ of the separations to the first period, and 138.6 to the second. This way we can complete the life table by computing the total years at risk for each period, and hence the hazard rate estimates.

In principle, we could do a second round of approximation, based on the updated hazard rate estimates. In fact, there is no real need to do this. The approximations are not very sensitive to the exact value of the hazard rate. If we do substitute the new values in, the estimated fraction of separations occurring in the first year will shift only from 0.418 to 0.419.

We are now in a position to use the model to draw some potentially interesting conclusions. For instance, we may be interested to know the probability that a cohabitation with children will end in separation. We need to decide what to do with the lack of observations after 5 years. For simplicity, let us assume that rates remain constant after that point, so that all cohabitations would eventually end in one of these fates. Applying the formula (9.1), we see that

$$\mathbb{P}\{\text{separate}\}=\int_0^\infty\mu_x^S\bar{F}(x)dx.$$

Table 9.4: Second approximation Multiple decrement life table.

x	ℓ_x	d_x^M	d_x^S	$\tilde{\ell}_x$	$\hat{\mu}_x^M$	$\hat{\mu}_x^S$
0–1	1000	180	99.4	860.2	0.209	0.116
1–3	720.6	120	138.6	1183	0.101	0.116
3–5	462	90	79	755	0.119	0.105

We have then

$$\begin{aligned}
\mathbb{P}\{\text{separate}\} &= 0.116 \int_0^1 e^{-0.325x} dx + 0.116 \int_1^3 e^{-0.325-0.217(x-1)} dx + 0.105 \int_3^\infty e^{-0.759-0.224(x-3)} dx \\
&= \frac{0.116}{0.325} \left[1 - e^{-0.325}\right] + \frac{0.116}{0.217} \left[e^{-0.325} - e^{-0.759}\right] + \frac{0.105}{0.224} \left[e^{-0.759}\right] \\
&= 0.099 + 0.136 + 0.219 \\
&= 0.454.
\end{aligned}$$

Lecture 10

Continuous-time Markov chains

10.1 General Markov chains

Let $\mathbb{S}$ be a countable state space, and $M = (M_n)_{n\geq 0}$ or $X = (X_t)_{t\geq 0}$ a time-homogeneous Markov chain with unknown transition probabilities/rates. In this lecture, we will develop methods to construct maximum likelihood estimators for the transition probabilities/ rates. You may think of a population model where birth rates and death rates may depend on the population size, and also multiple births (or maybe immigration) and multiple deaths (accidents, disasters, emigration) are allowed.

10.1.1 Discrete time, estimation of Π-matrix

Suppose we start a Markov chain at $M_0 = i_0$ and then observe $(M_0, \ldots, M_n) = (i_0, \ldots, i_n)$. The general transition matrix $\Pi = (\pi_{ij})_{i,j\in\mathbb{S}}$ contains our parameters π_{ij}, $i, j \in \mathbb{S}$, and we can write down the likelihood (probability mass function) for our observations

$$\prod_{k=1}^{n} \pi_{i_{k-1},i_k} = \prod_{i\in\mathbb{S}} \prod_{j\in\mathbb{S}} \pi_{ij}^{n_{ij}} \tag{10.1}$$

where n_{ij} is the number of transitions from i to j. Now note that $p_{ii} = 1 - \sum_{j\neq i} p_{ij}$ so that, as before, the maximum likelihood estimators are

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i}, \qquad \text{provided } n_i = \sum_{j\in\mathbb{S}} n_{ij} = \#\{0 \leq k \leq n-1 : i_k = i\} > 0. \tag{10.2}$$

If $n_i = 0$, then the likelihood is the same for all π_{ij}, $j \in \mathbb{S}$.

10.1.2 Estimation of the Q-matrix

Suppose we start a continuous-time Markov chain at $X_0 = i_0$ and observe $(X_s)_{0\leq s\leq T_n}$ where T_n is the nth transition time and record the data as successive holding times $T_j - T_{j-1} = h_{j-1}$ and sequence of states of the jump chain $(i_0, \ldots, i_n)$. The general Q-matrix $(q_{ij})_{i,j\in\mathbb{S}}$ contains the parameters q_{ij}, $i, j \in \mathbb{S}$, and the likelihood, as a product of densities of holding times and transition probabilities, is given by

$$\prod_{k=1}^{n} \lambda_{i_{k-1}} \exp\{-\lambda_{i_{k-1}} h_{j-1}\} \frac{q_{ik-1,i_k}}{\lambda_{i_{k-1}}} = \prod_{i\in\mathbb{S}} \prod_{j\neq i} q_{ij}^{n_{ij}} \exp\{-e_i q_{ij}\}, \tag{10.3}$$

where $\lambda_i = -q_{ii} = \sum_{j \neq i} q_{ij}$, n_{ij} is the number of transitions from i to j and e_i is the total time spent in i. This is maximised factor by factor by

$$\hat{q}_{ij} = \hat{q}_{ij}(i_0, h_0, i_1, \ldots, h_{n-1}, i_n) = \frac{n_{ij}}{e_i}, \qquad i \neq j. \tag{10.4}$$

In fact, the holding times and transitions may come from several chains (with the same unknown Q-matrix) without affecting the form of the likelihood, if we define

$$n_{ij} = n_{ij}^{(1)} + \ldots + n_{ij}^{(r)} \qquad \text{and} \qquad e_i = e_i^{(1)} + \ldots + e_i^{(r)} \tag{10.5}$$

for observed chains $(X_s^{(k)})_{0 \leq s \leq T_{n_k}^{(k)}}$, $1 \leq k \leq r$. This is useful to save time by simultaneous observation, and to reach areas of the state space not previously visited (e.g. for reducible or transient chains).

10.2 The induced Poisson process

In order to derive a rigorous estimation theory for a general finite-state Markov process, we consider the embedded Poisson processes that comes from considering the process only when it is in a given state. You might imagine a "state-x estimator" that is tasked with estimating the transition rates to all other states from state x; its clock runs while the process is in state x, and it counts the transitions, but it slumbers when the process is in any other state.

Suppose we run infinitely many copies of the Markov process, started in states $X_0^{(1)}, X_0^{(2)}, \ldots$ for some lengths of time $S^{(1)}, S^{(2)}, \ldots$. Suppose that the times $S^{(i)}$ are stopping times: That is, they do not depend upon knowing the future of the process. (For a precise technical definition, see any basic text on stochastic processes, such as [KT81].) The realisation i makes transitions at times $T_1^{(i)}, \ldots, T_{N_i}^{(i)}$ to states $X_1^{(i)}, \ldots, X_{N_i}^{(i)}$. (We do not wish to rule out the simple possibility that there is only one run of a positive recurrent Markov process. To admit that alternative, though, would complicate the notation. Instead, we may suppose that the single infinite run is broken into infinitely many pieces, for instance by stopping after each full unit of time, and restarting with $X_0^{(i+1)} = X_{N_i}^{(i)}$.) We assume that the realisations are independent, except that the starting state of a realisation may be dependent on

Consider some fixed state x. Suppose all K realisations visit state x a total of M_x times, and let $\tau_x(j)$ be the length of the j-th sojourn in x — so that $E_x := \tau_x(1) + \cdots \tau_x(M_x)$ is the total of all the time intervals when the process is in state x. Of the sojourns in x, some end with a transition to a new state, and some end because a stopping time $S^{(i)}$ intervened; define

$$\delta_x(j) = \begin{cases} 1 & \text{if sojourn } j \text{ ends with a transition;} \\ 0 & \text{if sojourn } j \text{ ends with a stopping time.} \end{cases}$$

Consider now the random interval $[0, E_x]$, and the set of points

$$\mathcal{S}_x := \big\{ \tau_x(1) + \cdots + \tau_x(j) : \text{ s.t. } \delta_x(j) = 1 \big\}.$$

The idea is that we take the events that occur only while the process is waiting at state x out, and stitch them together. Theorem 2 tells us that we obtain thereby a Poisson process. An illustration may be found in Figure 10.1. We start with a strong restatement of the "memoryless" property of the exponential distribution.

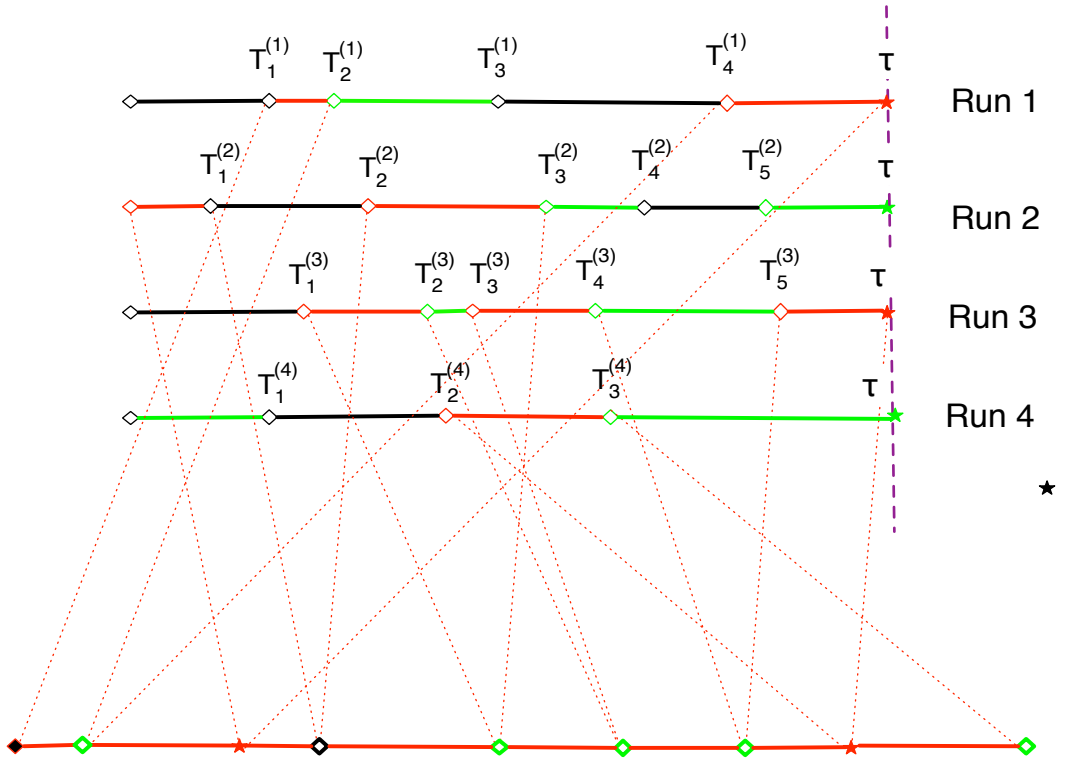


Figure 10.1: Illustration of the "stitching-together" construction, by which the process confined to a particular state generates a marked Poisson process. The Markov process has three states, represented by red, green, and black. We are estimating the transition rates from the red state. The diamond shapes represent the colour to which transitions are made. Stars represent censored observations; that is, times at which a realisation of the process was ended (at the time τ) in state R, without having transitioned out. The estimates based on these observations would be $\hat{q}_{RG} = 5/E$ and $\hat{q}_{RB} = 1/E$, where E is the total length of the red line at the bottom.

Lemma 1. *Suppose $T_1, T_2, \ldots$ is an i.i.d. sequence of exponential random variables with parameter λ, and $S_1, S_2, \ldots$ independent random variables such that each S_i is a stopping time with respect to T_i. That is, $T_i - t$ is independent of S_i on the event $\{S_i \le t\}$, for any fixed t. Let $K = \min\{k : T_k \le S_k\}$. Then*

$$T_* := T_K + \sum_{i=1}^{K-1} S_i$$

is exponential with parameter λ.

Proof. The stopping-time property tells us that $(T_i - S_i)$ is independent of S_i on the event $\{T_i > S_i\}$. Consequently, conditioned on $\{T_i > S_i = s\}$, $(T_i - S_i)$ has the distribution of $(T_i - s)$ conditioned on $\{T_i > s\}$, which is exponential with parameter λ; and conditioned on $\{T_i \le s\}$, T_i has exponential distribution with parameter λ. Conditioned on $\{K = 1\}$, then, it is immediately true that $T_* = T_1$ has the correct distribution. Suppose now that T_* has the correct distribution, conditioned on $\{K = k\}$. Then conditioned on $\{K = k+1\}$, $T_* := T_{k+1} + S_k + \sum_{i=1}^{k-1} S_i$. Note that $S_k + T_{k+1}$ conditioned on $\{K = k+1\}$ has the same distribution as T_k conditioned on $\{K = k\}$ (by the induction hypothesis). Since either of these is independent of $\sum_{i=1}^{k-1} S_i$, the distribution is the same T_* conditioned on $\{K = k+1\}$ is identical to the distribution conditioned on $\{K = k\}$, which completes the induction. □

Theorem 2. *The random set $\mathcal{S}_x$ is a Poisson process with rate $q_x := \sum_{y \ne x} q_{xy}$, and the total time E_x is a stopping time for the process. If we condition on $(E_x : x \in \mathcal{X})$, the processes corresponding to different states are independent. Finally, conditioned on $\mathcal{S}_x$, the transitions that take place at the times $\mathcal{S}_x$ are independent, with the probability of transitioning to y being q_{xy}/q_x.*

Proof. Consider the interarrival time between two points of $\mathcal{S}_x$. By Lemma 1 it is exponential with parameter q_x. By the Markov property, the interarrival times are all independent. Hence, these are independent Poisson processes. The independence of the transitions from the waiting times is standard. □

Define $N_x(s)$ to be the number of visits to state x up to total time s (where total time is measured by stitching together the processes $X_\cdot^{(1)}$, followed by $X_\cdot^{(2)}$, and so on) which end in a transition (as opposed to ending in a stopping time, and shift to a new realisation of the process). Let $N_{xy}(s)$ be the number of visits to state x up to total time s which end in a transition to state y. Let $E_x(s)$ be the total amount of time spent in state x up to total time s. (Thus, $\sum_{x \in \mathcal{X}} E_x(s) = s$ identically.)

Consequences of Theorem 2 are:

MLE The maximum likelihood estimator for the rate of a Poisson process is # events/total time. Thus, if we observe realisations of the process which add up to total time S (where S may be a random stopping time), the MLE for q_{xy} is

$$\hat{q}_{xy}(S) = \frac{N_{xy}(S)}{E_x(S)}. \tag{10.6}$$

Consistency $\lim_{s\to\infty} \frac{N_{xy}(S)}{E_x(S)} = q_{xy}$, on the event $\{\lim_{s\to\infty} E_x(s) = \infty\}$. (Question to consider: How would it be possible to arrange the realisations of the process so that the condition $\{\lim_{s\to\infty} E_x(s) = \infty\}$ does not have probability 1?)

Sampling dist. Suppose we run realisations of the process until a random time S that we will call $S_x(t) := \inf\{s : E_x(s) = t\}$; that is, we run the process (in its various successive realisations) until such time as the total time spent in x is exactly t. Then the estimator $\hat{q}_{xy}(S)$ is equal to $N_{xy}(S)/E_x(S) = N_{xy}(S)/t$. Since $N_{xy}(S)$ has Poisson distribution with parameter tq_{xy}, this tells us the distribution of $\hat{q}_{xy}(S)$. Its expectation is q_{xy}, and its variance is q_{xy}/t. As $t \to \infty$, $\hat{q}_{xy}(S_x(t))$ converges to a normal distribution.

The sampling distribution is a complicated matter. If we have observed up to time $S_x(t)$ then we know the exact distribution of $\hat{q}_{xy}$, and we can compute approximate $100\alpha\%$ confidence intervals as $\hat{q}_{xy} \pm z_{(1-\alpha/2}\sqrt{\hat{q}_{xy}/t}$, where z is the appropriate quantile of the normal distribution. Even then, we do not have the exact distribution even for the estimators of transition rates starting from any other state.

Can this be a serious problem? Suppose we observe instead up to a constant total time s. Is the distribution substantially different? In some respects it is, particularly in the tails. Suppose we decompose the estimate by the number of visits to x.

$$\hat{q}_{xy}(s) = \frac{N_{xy}(s)}{E_x(s)} = \sum_{n=0}^{\infty} \mathbf{1}_{\{N_x(s)=n\}} \frac{N_{xy}(s)}{E_x(s)}.$$

The summand corresponding to $n = 0$ is $0/0$, which is problematic. Moving on to $n = 1$, we have with probability q_{xy}/q_x the expression $1/E$, where E is exponential with parameter q_x. This has expectation ∞. Consequently, $\hat{q}_{xy}(s)$ also has infinite expectation. Other choices of a random time S at which to observe the process can similarly distort the distribution.

On the other hand, this is only a problem with the expectation and variance, not with the main bulk of the distribution. That is, as long as S is chosen so that there will be, with high probability, a large number of visits to x, the normal approximation should be fairly accurate.

The general rule for estimating transition rates is

$$\hat{q}_{xy} = \frac{\#\text{ transitions } x \to y}{\text{total time spent in state } x}$$

$$\mathrm{Var}(\hat{q}_{xy}) \approx \frac{\hat{q}_{xy}}{\text{total time spent in state } x}$$

We compute an approximate $100\alpha\%$ confidence interval as $\hat{q}_{xy} \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{q}_{xy})}$.
If the number of transitions is not very large, we may do better estimating $\hat{q}_{xy}$ from the Poisson parameter estimated by N_{xy}. Exact confidence intervals may be computed, using the identity

$$\mathbb{P}\big\{k \le N_s\big\} = \mathbb{P}\big\{T_k \le s\big\} = \mathbb{P}\big\{2k\mu T_k \le 2k\mu s\big\} = \alpha \qquad \text{for } \mu = \chi_{2k,\alpha}/(2ks).$$

This is carried out in the exercises.

10.3 Parametric and time-dependent models

So far, we have assumed that the Q-matrix was completely arbitrary, i.e. with entries $q_{ij} \ge 0$, $j \ne i$, and $q_{ii} = -\sum_{j:j\ne i} q_{ij} > -\infty$. We estimated all "parameters" q_{ij} by observing a Markov

chain with that unknown Q-matrix (or several independent Markov chains with the same unknown Q-matrix).

On the other hand, we studied the multiple decrement model, which we can view as a continuous-time Markov chain, where the Q-matrix contains lots of zeros, namely everywhere except in the first row. Here, the maximum likelihood estimates that we derived were of the same form

$$\frac{\text{numbers of transitions from } i \text{ to } j}{\text{total time spent in } i}, \tag{10.7}$$

except that we did not specify the zero rows as estimates but as model assumptions.

Here, we will merge ideas and study more systematically Markov chain models where the Q-matrices are not completely unknown. Instead, we assume/know some structure, e.g. certain zero entries and/or that certain transition rates are the same or stand in a known relationship to each other.

Secondly, we will incorporate time-dependent transition rates (as we had in the multiple decrement model) into the general Markov model.

10.3.1 Example: Marital status model

A natural model for marital status consists of five states "bachelor" (B), "married" (M), "widowed" (W), "divorced" (D) and "dead" (Δ).

We can set up a model with 9 parameters corresponding to the 9 possible transitions ($B \to M$, $B \to \Delta$, $M \to W$, $M \to D$, $M \to \Delta$, $W \to M$, $W \to \Delta$, $D \to M$, $D \to \Delta$). Note that also state Δ is absorbing and there is no reason to continue to observe chains that have run into this state. This means that we agree that four entries vanish:

$$q_{\Delta B} = q_{\Delta M} = q_{\Delta W} = q_{\Delta D} = 0. \tag{10.8}$$

Furthermore, it is also impossible to have direct transitions between B, W and D or indeed to go from M to B, so we also know in advance that

$$q_{BD} = q_{DB} = q_{BW} = q_{WB} = q_{DW} = q_{WD} = q_{MB} = 0. \tag{10.9}$$

With states arranged in the above order, this gives a Q-matrix

$$Q = \begin{pmatrix} -\alpha - \mu_B & \alpha & 0 & 0 & \mu_B \\ 0 & -\nu - \delta - \mu_M & \nu & \delta & \mu_M \\ 0 & \sigma & -\sigma - \mu_W & 0 & \mu_W \\ 0 & \rho & 0 & -\rho - \mu_D & \mu_D \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{10.10}$$

Alternatively, we can *assume* that the death rate does not depend on the current state B, M, W or D, so that the Q-matrix only contains 6 parameters as

$$Q = \begin{pmatrix} -\alpha - \mu & \alpha & 0 & 0 & \mu \\ 0 & -\nu - \delta - \mu & \nu & \delta & \mu \\ 0 & \sigma & -\sigma - \mu & 0 & \mu \\ 0 & \rho & 0 & -\rho - \mu & \mu \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{10.11}$$

Finally, we can allow age-varying transition rates by having $Q(t)$, where now the parameters $\alpha(t)$, $\mu(t)$, $\nu(t)$, $\delta(t)$, $\sigma(t)$ and $\rho(t)$ depend on age t.

10.3.2 The general simple birth-and-death process

The general simple birth-and-death process is a continuous-time Markov chain with Q-matrix

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & \cdots \\ \mu_1 & -\lambda_1 - \mu_1 & \lambda_1 & 0 & 0 & \cdots \\ 0 & \mu_2 & -\lambda_2 - \mu_2 & \lambda_2 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{10.12}$$

where birth rates λ_j and death rates μ_j are in arbitrary (unknown) dependence on j. Note that this model has an infinite number of parameters, but just as with unspecified maximal ages in the single decrement model, we would only estimate $(\lambda_0, \mu_1, \ldots, \mu_{\max}, \lambda_{\max})$, where indeed $\lambda_{\max}$ may be left unspecified or equal to zero for the highest observed population size.

This model has the same maximum likelihood estimators as the general Q-matrix with all entries unknown, since that likelihood factorises completely, and given samples from simple birth-and-death processes, there are no multiple births and deaths, so the maximum likelihood estimates of the multiple birth and death rates (jumps of sizes two or higher) are zero. The (remaining) likelihood is

$$\prod_{k=1}^{n} q_{X_{T_{k-1}}, X_{T_k}} \exp\{-\sum_{j: j \neq X_{T_{k-1}}} q_{X_{T_{k-1}}, j}(T_k - T_{k-1})\} = \prod_{i \in \mathbb{N}} \prod_{|j-i|=1} q_{ij}^{N_{ij}} \exp\{-E_i q_{ij}\}, \tag{10.13}$$

where now $q_{i,i+1} = \lambda_i$ and $q_{i,i-1} = \mu_i$, $i \geq 1$. N_{ij} is the number of transitions from i to j and E_i is the total time spent in i. Note that we have taken the liberty and written the likelihood in terms of the underlying random variables, not the realisations.

We note a general phenomenon: if transitions are impossible, and, in particular there are no observations of such transitions in the samples, their likelihood contribution is maximised by a zero rate. Note that vice versa, a given sample may not contain some other transitions although they are possible. In this case, the estimate of the corresponding transition rate is zero, but not usually the estimator, which is non-zero with positive probability for all transitions that are possible within the given number of steps from the given initial values.

10.3.3 Lower-dimensional parametric models of simple birth-and-death processes

Often population models come with some additional structure. The simplest structure is that of independent individuals each giving birth repeatedly at rate λ until their death at rate μ. Here $\lambda_j = j\lambda$ and $\mu_j = j\mu$, $j \in \mathbb{N}$, are all expressed in terms of two parameters λ and μ. The likelihood in this model is the same as in the general model, but has to be factorised as

$$\prod_{i \in \mathbb{N}} \prod_{|j-i|=1} q_{ij}^{N_{ij}} \exp\{-E_i q_{ij}\} = \left(\prod_{i \in \mathbb{N}} (i\lambda)^{N_{i,i+1}} \exp\{-E_i i\lambda\} \right) \left(\prod_{i \in \mathbb{N}} (i\mu)^{N_{i,i-1}} \exp\{-E_i i\mu\} \right) \tag{10.14}$$

to separate the two parameters. This can best be maximised via the log likelihood, which for the μ-factor is

$$\sum_{i=1}^{\infty} \left(N_{i,i-1}(\log(i) + \log(\mu)) - E_i i\mu \right). \tag{10.15}$$

Differentiation leads to the maximum likelihood estimator $\hat{\mu} = D/W$ where $D = \sum_i N_{i,i-1}$ is the total number of deaths and $W = \sum_i iE_i$ is the weighted sum of exposure times at population sizes $i \in \mathbb{N}$. This quantity has again the interpretation as the total time spent at risk since in state i there are i individuals at risk to die. Similarly, $\hat{\lambda} = B/W$, where B is the total number of births. (W is also the total time spent at risk to give birth.)

Note that the state 0 is absorbing, so an nth transition may never come. This can be helped by running several Markov chains. The likelihood function of the given sample is the one given, if we understand that N_{ij} are aggregate counts and E_i are aggregate waiting times, $i, j \in \mathbb{N}$.

10.4 Time-varying transition rates

10.4.1 Maximum likelihood estimation

We take up the setting of a general Markov model $(X_t)_{t\geq 0}$, but have the Q-matrix depend on t, $Q(t)$. For simplicity think of t as age. Denote the finite or countably infinite state space by $\mathbb{S}$. Given we have reached state $i \in \mathbb{S}$ aged exactly $y \in [0, \infty)$, the situation is exactly as for the multiple decrement model, there are competing hazards $q_{ij}(y + t)$, $t \geq 0$, $j \neq i$, and the total holding time in state i has hazard rate

$$-q_{ii}(y + t) = \sum_{j:j\neq i} q_{ij}(y + t), \qquad t \geq 0. \tag{10.16}$$

Given the holding time is $Z = t$, the transition is from i to j with probability

$$\mathbb{P}(X_{y+Z} = j | X_y = i, Z = t) = \frac{q_{ij}(y + t)}{\sum_{j:j\neq i} q_{ij}(y + t)}. \tag{10.17}$$

To be able to estimate time-varying transition rates, we require more than one realisation of X, say realisations $(X_t^{(m)})_{0\leq t\leq T_{n_m}^{(m)}}$, $m = 1, \ldots, r$, where T_{n_m} is the n_mth transition time of $X^{(m)}$. Then the likelihood is given by

$$\prod_{m=1}^{r} \left(\prod_{k=1}^{n_m} q_{X^{(m)}_{T^{(m)}_{k-1}}, X^{(m)}_{T^{(m)}_k}}(T_k^{(m)}) \exp\left\{ - \int_{T_{k-1}^{(m)}}^{T_k^{(m)}} \sum_{j:j\neq X^{(m)}_{T^{(m)}_{k-1}}} q_{X^{(m)}_{T^{(m)}_{k-1}}, j}(s) ds \right\} \right) \tag{10.18}$$

If we also postulate simplifying assumptions such as piecewise constant transition rates $q_{ij}(t) = q_{ij}(x + \frac{1}{2})$, $x \leq t < x + 1$, $x \in \mathbb{N}$, we can reexpress this in a factorised form as

$$\prod_{x\in\mathbb{N}} \prod_{i\in\mathbb{S}} \prod_{j\neq i} q_{ij}(x + \tfrac{1}{2})^{N_{ij}(x)} \exp\left\{ -E_i(x) q_{ij}(x + \tfrac{1}{2}) \right\}, \tag{10.19}$$

where $N_{ij}(x)$ is the number of transitions from i to j at an age x, i.e. aged t with $x \leq t < x + 1$, and $E_i(x)$ is the total time spent in state i while aged x.

We read off the maximum likelihood estimators for all $x \in \mathbb{N}$ and $i \in \mathbb{N}$ with $E_i(x) > 0$:

$$\hat{q}_{ij}(x + \tfrac{1}{2}) = \frac{N_{ij}(x)}{E_i(x)}. \tag{10.20}$$

10.4.2 Example

Clearly, a reasonably complete set of reasonably reliable estimates can only be obtained if the state space $\mathbb{S}$ is small and the number of observations is very large, e.g., in the illness-death model with three states H=able, S=sick and D=dead, with age-dependent sickness rates σ_x from H to S, recovery rates ρ_x from S to H and death rates δ_x from H to D and γ_x from S to D.

Suppose, we observe r individuals over their whole life $[0, \tau_d^{(m)}]$, then we get estimates

$$\hat{\delta}_{x+\frac{1}{2}} = \frac{d_x}{v_x}, \quad \hat{\gamma}_{x+\frac{1}{2}} = \frac{c_x}{w_x}, \quad \hat{\sigma}_{x+\frac{1}{2}} = \frac{s_x}{v_x}, \quad \hat{\rho}_{x+\frac{1}{2}} = \frac{r_x}{w_x}, \tag{10.21}$$

where v_x (w_x) is the total waiting time of lives aged x in the able (ill) state, and d_x, c_x, s_x, r_x are the aggregate counts of the respective transitions at age x.

10.4.3 Construction of the stochastic process $(X_t)_{t\geq 0}$

This section is *non-examinable* and deals with the Probability behind minimal $(\nu, (Q(t))_{t\geq 0})$ Markov chains, where ν is an initial distribution on $\mathbb{S}$ and $(Q(t))_{t\geq 0}$ is a time-dependent Q-matrix.

We first give a construction analogous to the maze construction for continuous-time Markov chains with constant transition rates, but note that the jump-chain holding description was rather vague, so we will not "prove" but only "indicate" why the process we construct does what we want. In fact, you may wish to take the maze construction as the definition of a minimal $(\nu, (Q(t))_{t\geq 0}$ Markov chain.

We construct counting processes $(N_t^{(ij)})_{t\geq 0}$ for all pairs $i, j \in \mathbb{S}$, $i \neq j$, independent. Fix i and j and consider a Poisson process $\tilde{N}^{(ij)}$ with unit rate. Then define

$$N_t^{(ij)} = \tilde{N}^{(ij)}_{\int_0^t q_{ij}(s)ds} \tag{10.22}$$

This is a time-inhomogeneous Poisson process. It is obvious that $N^{(ij)}$ is still a counting process since the jumps of N and $\tilde{N}$ are in $1-1$ correspondence

$$N_t^{(ij)} - N_{t-}^{(ij)} = \tilde{N}^{(ij)}_{\int_0^t q_{ij}(s)ds} - \tilde{N}^{(ij)}_{\int_0^t q_{ij}(s)ds-}. \tag{10.23}$$

N still has independent increments with Poisson distributions, since

$$\begin{aligned} N_{t_n}^{(ij)} - N_{t_{n-1}}^{(ij)} = \tilde{N}^{(ij)}_{\int_0^{t_n} q_{ij}(s)ds} - \tilde{N}^{(ij)}_{\int_0^{t_{n-1}} q_{ij}(s)ds} &\sim \text{Poi}\left(\int_0^{t_n} q_{ij}(s)ds - \int_0^{t_{n-1}} q_{ij}(s)ds\right) \\ &= \text{Poi}\left(\int_{t_{n-1}}^{t_n} q_{ij}(s)ds\right), \end{aligned}$$

but note that these increments are no longer stationary for all increment lengths, unless $q_{ij} \equiv q_{ij}(s)$ does not depend on s, in which case $N^{(ij)}$ is simply a (homogeneous) Poisson process with rate q_{ij}.

Next we define aggregate processes

$$N_t^{(i)} = \sum_{j\neq i} N_t^{(ij)} \sim \text{Poi}\left(\int_0^t \lambda_i(s)ds\right), \qquad t \geq 0, i \in \mathbb{S},$$

also inhomogeneous Poisson processes. Note that the first jump time $T_1^{(i)}$ of $N^{(i)}$ has survival function

$$\mathbb{P}(T_1^{(i)} > t) = \mathbb{P}(N_t^{(i)} = 0) = \exp\left\{-\int_0^t \lambda_i(s)ds\right\},$$

i.e. hazard rate $\lambda_i(t)$. Also $T_1^{(i)} = \inf\{T_1^{(ij)}, j \neq i\}$ is as for the multiple decrement model. Similarly, for $T_1^{(i)}(x) = \inf\{t \geq x : N_t^{(i)} \neq N_x^{(i)}\}$, we have

$$\begin{aligned}\mathbb{P}\left(T_1^{(i)}(x) > x + t\right) = \mathbb{P}\left(N_{x+t}^{(i)} - N_x^{(i)} = 0\right) &= \exp\left\{-\int_x^{x+t} \lambda_i(s)ds\right\} \\ &= \exp\left\{-\int_0^t \lambda_i(x+s)ds\right\},\end{aligned}$$

and this identifies a hazard rate of $\lambda_i(x+t)$. Furthermore, as calculated for $T = \min\{T_j : 1 \leq j \leq m\}$ in the multiple decrement model, we have for $T_1^{(i)}(x) = \inf\{T_1^{(ij)}(x), j \neq i\}$

$$\mathbb{P}\left(T_1^{(i)}(x) = T_1^{(ij)}(x) \middle| T_1^{(i)}(x) = t\right) = \frac{q_{ij}(x+t)}{\lambda_i(x+t)}.$$

The construction is as follows. Take $M_0 \sim \nu$, independent from all Poisson processes $(N_t^{(ij)})_{t\geq 0}$, $T_0 = 0$, and define inductively jump times

$$T_{n+1} = \inf\{t > T_n : N_t^{(M_n)} \neq N_{T_n}^{(M_n)}\},$$

and jump destinations

$$M_{n+1} = j \qquad \Longleftrightarrow \qquad N_{T_{n+1}}^{(M_n,j)} \neq N_{T_{n+1}-}^{(M_n,j)},$$

$n \in \mathbb{N}$. Then specify X as

$$X_t = M_n \qquad \Longleftrightarrow \qquad T_n \leq t < T_{n+1}.$$

In general, M is not a Markov chain, and holding times $T_{n+1} - T_n$ are not conditionally independent given M, but X has been constructed from independent Poisson processes only, as in the constant-Q case, and it can be shown that X has a Markov property, that we formulate below.

First note that we can, more generally, construct a $(\nu, x, (Q(t))_{t\geq 0}$ chain $(X_t)_{t\geq x}$ starting at time x (rather than 0) from an initial distribution $X_x \sim \nu$, simply by changing $T_0 := 0$ to $T_0 := x$, while keeping the remainder of the construction.

The Markov property of X now states that $(X_{x+t})_{t\geq 0}$ is conditionally independent of $(X_r)_{0\leq r\leq x}$ given $X_x = i$. Given $X_x = i$, the post-x process is a $(\delta_i, x, (Q(t))_{t\geq 0})$ Markov chain, where $\delta_i = (\delta_{ij})_{j\in\mathbb{S}}$ is the Dirac probability mass function putting all mass in i, $\delta_{ii} = 1$. This Markov property is again a consequence of the maze construction, since the post-x process only depends on the current state and the (inhomogeneous, but independent-increment) Poisson processes after time x.

We can then derive, under some further regularity conditions, as for the constant-rate case, an infinitesimal description,

$$\mathbb{P}(X_{x+t} = j | X_x = i) = \mathbb{P}(T_1^{(i)}(x) \leq x + t, T_1^{(ij)}(x) = T_1^{(i)}(x)) + o(t) = q_{ij}(x)t + o(t), \quad (10.24)$$

for $i \neq j$, as $t \downarrow 0$, and forward and backward equations

$$\frac{\partial}{\partial t}P(s,t) = P(s,t)Q(t) \qquad \text{and} \qquad \frac{\partial}{\partial s}P(s,t) = Q(s)P(s,t), \quad (10.25)$$

for the transition matrices $P(s,t) = (p_{ij}(s,t))_{i,j\in\mathbb{S}}$, where $p_{ij}(s,t) = \mathbb{P}(X_t = j | X_s = i)$.

10.5 Occupation times

The last topic we consider is the generalisation of the formulas that we had earlier for life expectancy. For a lifetime with constant mortality rate μ, the expected lifetime is μ^{-1}. Consider now the illness model described in section 10.4.2

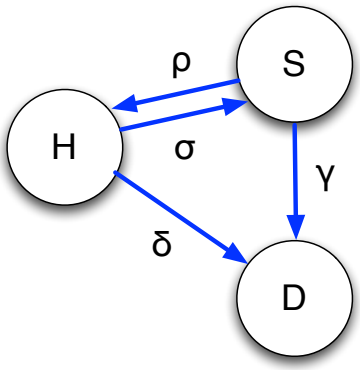


Figure 10.2: Diagram of the Illness model

We say that a matrix is **negative-definite** if all of its eigenvalues are negative. Define

$$T_x(t) := \text{total time spent in state } x \text{ up to time } t,$$

and ${}_yE_x(t) := \mathbb{E}_y[T_x(t)]$, where $\mathbb{E}_y$ represents the expectation given that the process starts in state y.

Theorem 3. *Let Q be the $(m+k)\times(m+k)$ transition matrix for a continuous-time discrete-space Markov process with m absorbing states and k non-absorbing states. Let Q_* be the $k\times k$ submatrix consisting of transition rates among the non-absorbing states. If Q_* is irreducible and some row has negative sum, then Q_* is negative definite, and the process is eventually absorbed with probability 1. Then*

$$ {}_yE_x(t) = Q_*^{-1}\left(e^{tQ_*} - I\right).$$

The limit ${}_yE_x := \lim_{t\to\infty} {}_yE_x(t)$ is finite and given by the (y,x) entry of $-Q_*^{-1}$.

Proof. The matrix of transition probabilities at time t is given by $P_t = e^{tQ}$. Then

$$\begin{aligned}
{}_yE_x(t) &= \mathbb{E}_y\left[\int_0^t \mathbf{1}_{\{X_s=x\}}ds\right] \\
&= \int_0^t P_s(y,x)ds \\
&= \left[\int_0^t E^{sQ}ds\right](y,x) \\
&= \left[\int_0^t E^{sQ_*}ds\right](y,x) \qquad \text{because the other states are absorbing} \\
&= Q_*^{-1}\left(e^{tQ_*} - I\right)(y,x).
\end{aligned}$$

By negative-definiteness of Q_* we have $\lim_{t\to\infty} e^{tQ_*} = 0$, so this converges to $-Q_*^{-1}$. □

We are also interested in the state that the process ends up in. We state the following result in terms of the diagonalisation of Q. Of course, in the special case where Q is not diagonalisable, we can carry out the same construction with the Jordan Normal Form.

Theorem 4. *Let $v_1, \ldots, v_m$ be the right-eigenvectors (columns) of Q with eigenvalue 0 (in other words, a basis for the kernel of Q), such that v_i has a 1 in coordinate $k+i$ and 0 for the remainder of the absorbing states. Then*

$$P_j\big\{\textit{absorbed in state } k+i\big\} = (v_i)_j.$$

Proof. Fix some i, and let $v_j = P_j\big\{\text{absorbed in state } k+i\big\}$. Let X_t be the position of the process at time t. Then $\mathbb{E}_{j'}[v_{X_t}] = P^t(j', \cdot)v$. Note that this function is constant in time (by the Chapman-Kolmogorov equation). By the Forward Equation, for all $t > 0$

$$0 = \frac{d}{dt} P^t(j', \cdot)v = P^t(j', \cdot)Qv,$$

which implies that $Qv = 0$, since $\lim_{t \downarrow 0} P^t$ is the identity matrix. Obviously v has the stated values on the absorbing states. □

10.5.1 The multiple decrements model

The simplest application of this formula is to the multiple decrements model. In that case, we have just a single non-absorbing state 0, and absorbing states $1, 2, \ldots, m$. Then $Q_* = (-\lambda_+)$, so that the expected time spent in state 0 is $1/\lambda_+$, which we already knew.

The only non-trivial eigenvector is

$$\begin{pmatrix} -1 & \frac{\lambda_1}{\lambda_+} & \cdots & \frac{\lambda_m}{\lambda_+} \end{pmatrix}.$$

Thus

$$R^{-1} = \begin{pmatrix} -1 & \frac{\lambda_1}{\lambda_+} & \frac{\lambda_2}{\lambda_+} & \cdots & \frac{\lambda_m}{\lambda_+} \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Thus, the probability of ending up in state i is λ_i/λ_+.

10.5.2 The illness model

Consider now the illness model, with $\sigma = 0.1$, $\delta = 0.1$, $\gamma = 0.5$, and $\rho = 0$. The generator (taking the states in the order H, S, D, is

$$Q = \begin{pmatrix} -0.2 & 0.1 & 0.1 \\ 0 & -0.5 & 0.5 \\ 0 & 0 & 0 \end{pmatrix} \qquad Q_* = \begin{pmatrix} -0.2 & 0.1 \\ 0 & -0.5 \end{pmatrix}$$

We calculate

$$Q_*^{-1} = \begin{pmatrix} -5 & -1 \\ 0 & -2 \end{pmatrix}$$

Thus, a sick individual survives on average 2 years. A healthy individual survives on average 6 years, of which 1 year, on average, is spent sick.

There is only one absorbing state. If we want to study the state that individuals died from (sick or healthy), one approach is to make two absorbing states, one corresponding to death after being healthy, the other after being sick. The core Q_* matrix stays the same, but now Q becomes

$$\begin{pmatrix} -0.2 & 0.1 & 0.1 & 0 \\ 0 & -0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The eigenvectors with eigenvalue 0 are

$$\begin{pmatrix} 0.5 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} 0.5 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

These tell us that someone who starts sick will with certainty die from the sick state (since there is no recovery in this model), while an initially healthy individual will have probability 1/2 of dying from the healthy or the sick state.

Suppose the recovery rate ρ now becomes 1. Then

$$Q = \begin{pmatrix} -0.2 & 0.1 & 0.1 & 0 \\ 1 & -1.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad Q_* = \begin{pmatrix} -0.2 & 0.1 \\ 1 & -1.5 \end{pmatrix} \qquad Q_*^{-1} = \begin{pmatrix} -7.5 & -.5 \\ -5 & -1 \end{pmatrix}$$

Thus, a healthy individual will now live, on average, 8 years, of which only 0.5 will be sick, and someone who is sick will have 6 years on average, with 1 of those sick.

The eigenvectors with eigenvalue 0 are now

$$\begin{pmatrix} 0.75 \\ 0.5 \\ 1 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} 0.25 \\ 0.5 \\ 0 \\ 1 \end{pmatrix}.$$

Thus we see that when starting from state H, the probability of transitioning to D from H has gone up to 3/4. Starting from S, the probability is now 1/2 of transitioning to D from S, and 1/2 of transitioning from H. This is consistent with the observation we make from the jump chain, that the healthy person transitions to sick or to dead with equal probabilities. Thus,

$$P_H(\text{last state H}) = \frac{1}{2} + \frac{1}{2}P_S(\text{last state H}) = \frac{1}{2} + \frac{1}{4}.$$

Lecture 11

Survival analysis: Introduction

Reading: Cox and Oakes, chapter 1, 4, section 11.6, CT4 Unit 6.3-6.5, Klein and Moeschberger sections 3.1–3.4, Chapter 4

11.1 [

Censoring and truncation]**Incomplete observations: Censoring and truncation**

We begin by considering simple analyses but we will lead up to and take a look at regression on explanatory factors, as in linear regression part A. The important difference between survival analysis and other statistical analyses which you have so far encountered is the presence of censoring. This actually renders the survival function of more importance in writing down the models.

Right censoring occurs when a subject leaves the study before an event occurs, or the study ends before the event has occurred. For example, we consider patients in a clinical trial to study the effect of treatments on stroke occurrence. The study ends after 5 years. Those patients who have had no strokes by the end of the year are censored. If the patient leaves the study at time t_e, then the event occurs in (t_e, ∞).

Left censoring is when the event of interest has already occurred before enrolment. This is very rarely encountered.

Truncation is deliberate and due to study design.

Right truncation occurs when the entire study population has already experienced the event of interest (for example: a historical survey of patients on a cancer registry).

Left truncation occurs when the subjects have been at risk before entering the study (for example: life insurance policy holders where the study starts on a fixed date, event of interest is age at death).

Generally we deal with **right censoring &** sometimes **left truncation**.

Two types of independent **right censoring**:

Type I: All subjects start and end the study at the same fixed time. All are censored at the end of the study. If individuals have different but fixed censoring times, this is called *progressive type I* censoring.

Type II: study ends when a fixed number of events amongst the subjects has occurred.

Type III or random censoring: Individuals drop out or are lost to followup at times that are random, rather than predetermined. We generally assume that the censoring times are independent of the event times, in which case there is no need to distinguish this from Type I.

Skeptical question: Why do we need special techniques to cope with incomplete observations? Aren't all observations incomplete? After all, we never see all possible samples from the distribution. If we did, we wouldn't need any sophisticated statistical analysis.

The point is that most of the basic techniques that you have learned assume that the observed values are interchangeable with the unobserved values. The fact that a value has been observed does not tell us anything about what the value is. In the case of censoring or truncation, there is dependence between the event of observation and the value that is observed. In right-censoring, for instance, the fact of observing a time implies that it occurred *before* the censoring time. The distribution of a time *conditioned on its being observed* is thus different from the distribution of the times that were censored.

There are different levels of independence, of course. In the case of random (type III) censoring, the censoring time itself **is** independent of the (potentially) observed time. In Type II censoring, the censoring time depends in a complicated way on all the observation times.

11.2 Likelihood and Censoring

If the censoring mechanism is *independent* of the event process, then we have an easy way of dealing with it.

Suppose that T is the time to event and that C is the time to the censoring event.

Assume that all subjects may have an event or be censored, say for subject i one of a pair of observations $\left(\widetilde{t}_i, \widetilde{c}_i\right)$ may be observed. Then since we observe the minimum time we would have the following expression for the likelihood (using independence)

$$L = \prod_{\widetilde{t}_i < \widetilde{c}_i} f(\widetilde{t}_i) S_C(\widetilde{t}_i) \prod_{\widetilde{c}_i < \widetilde{t}_i} S(\widetilde{c}_i) f_C(\widetilde{c}_i)$$

Now define the following random variable:

$$\delta = \begin{cases} 1 & \text{if } T < C \\ 0 & \text{if } T > C \end{cases}$$

For each subject we observe $t_i = \min\left(\widetilde{t}_i, \widetilde{c}_i\right)$ and δ_i, observations from a continuous random variable and a binary random variable. In terms of these L becomes

$$L = \prod_i h(t_i)^{\delta_i} S(t_i) \prod_i h_C(t_i)^{1-\delta_i} S_C(t_i)$$

where we have used density = hazard $\times$ survival function.

NB If the censoring mechanism is independent (sometimes called non-informative) then we can ignore the second product on the right as it gives us no information about the event time. In the remainder of the course we will assume that the censoring mechanism is independent.

11.3 Data

Demographic v. trial data

Our models include a "time" parameter, whose interpretation can vary. First of all, in population-level models (for instance, a birth-death model of population growth, where the state represents the number of individuals) the time is true calendar time, while in individual-level models (such as our multiple-decrement model of death due to competing risks, or the healthy-sick-dead process, where there is a single model run for each individual) the time parameter is more likely to represent individual age. Within the individual category, the time to event can

literally be the age, for instance in a life insurance policy. In a clinical trial it will more typically be time from admission to the trial.

For example, consider the following data from a Sydney hospital pilot study, concerning the treatment of bladder cancer:

Time to cancer	Time to recurrence	Time between	Recurrence status
0.000	4.967	4.967	1
21.020	22.993	1.974	1
45.033	61.086	16.053	0
52.171	55.033	2.862	1
48.059	65.033	16.974	0

All times are in months. Each patient has their own zero time, the time at which the patient entered the study (accrual time). For each patient we record time to event of interest or censoring time, whichever is the smaller, and the status, $\delta = 1$ if the event occurs and $\delta = 0$ if the patient is censored. If it is the recurrence that is of interest, so in fact the relevant time, the "time between", is measured relative to the zero time that is the onset of cancer.

11.4 Non-parametric survial estimation

11.4.1 Review of basic concepts

Consider random variables $X_1, \ldots, X_n$ which represent independent observations from a distribution with cdf F. Given a class $\mathfrak{F}$ of possibilities for F, an **estimator** is a choice of the "best", on the basis of the data. That is, it is a function from $\mathbb{R}_+^n$ to $\mathfrak{F}$ which maps an collection of observations $(x_1, \ldots, x_n)$ to $\mathfrak{F}$.

Estimators for distribution functions may be either **parametric** or **non-parametric**, depending on the nature of the class $\mathfrak{F}$. The distinction is not always clear-cut. A parametric estimator is one for which the class $\mathfrak{F}$ depends on some collection of parameters. For example, it might be the two-dimensional family of all gamma distributions. A non-parametric estimator is one that does not impose any such parametric assumptions, but allows the data to "speak for themselves". There are intermediate non-parametric approaches as well, where an element of $\mathfrak{F}$ is not defined by any small number of parameters, but is still subject to some constraint. For example, $\mathfrak{F}$ might be the class of distributions with smooth hazard rate, or it might be the class of log-concave distribution functions (equivalent to having increasing hazard rate). We will also be concerned with **semi-parametric** estimators, where an underlying infinite-dimensional class of distributions is modified by one or two parameters of special interest.

The disadvantage of parametrisation is always that it distorts the observations; the advantage is that it allows the data from different observations to be combined into a single parameter estimate. (Of course, if the data are *known* to come from some distribution in the parametric family, the "distortion" is also an advantage, because the real distortion was in the data, due to random sampling.)

We start by considering nonparametric estimators of the cdf. These have the advantage of limiting the assumptions imposed upon the data, but the disadvantage of being too strictly limited by the data. That is, taken literally, the estimator we obtain from a sample of observed times will imply that only exactly those times actually observed are possible.

If there are observations $x_1, \ldots, x_n$ from a random sample then we define the empirical distribution function

$$\widehat{F}(x) = \frac{1}{n} \# \{x_i : x_i \leq x\}$$

This is an appropriate non-parametric estimator for the cdf if no censoring occurs. However if censoring occurs this has to be taken into account.

We measure the pair (X, δ) where $X = \min(T, C)$ and δ is as before

$$\delta = \begin{cases} 1 & \text{if } T < C \\ 0 & \text{if } T > C \end{cases}$$

Suppose that the observations are (x_i, δ_i) for $i = 1, 2 \ldots, n$.

$$\begin{aligned} L &= \prod_i f(x_i)^{\delta_i} S(x_i)^{1-\delta_i} \\ &= \prod_i f(x_i)^{\delta_i} \left(1 - F(x_i)\right)^{1-\delta_i} \end{aligned}$$

What follows is a heuristic argument allowing us to find an estimator for S, the survival function, which in the likelihood sense is the best that we can do. Notice first that there is no MLE if we model the failure time as a continuous random variable. Suppose T has density f, with survival function $S = 1 - F$

Suppose that there are failure times $(t_0 = 0 <)t_1 < \ldots < t_i < \ldots$. Let $s_{i1}, s_{i2}, \cdots, s_{ic_i}$ be the censoring times within the interval $[t_i, t_{i+1})$ and suppose that there are d_i failures at time t_i (allowing for tied failure times). Then the likelihood function becomes

$$\begin{aligned} L &= \prod_{fail} f(t_i)^{d_i} \prod_i \left(\prod_{k=1}^{c_i} \left(1 - F(s_{ik})\right) \right) \\ &= \prod_{fail} \left(F(t_i) - F(t_i-)\right)^{d_i} \prod_i \left(\prod_{k=1}^{c_i} \left(1 - F(s_{ik})\right) \right) \end{aligned}$$

where we write $f(t_i) = F(t_i) - F(t_i-)$, the difference in the cdf at time t_i and the cdf immediately before it.

Since $F(t_i)$ is an increasing function, and *assuming that it takes fixed values at the failure time points*, we make $F(t_i-)$ and $F(s_{ik})$ as small as possible in order to maximise the likelihood. That means we take $F(t_i-) = F(t_{i-1})$ and $F(s_{ik}) = F(t_i)$.

This maximises L by considering the cdf $F(t)$ to be a step function and therefore to come from a discrete distrbution, with failure times as the actual failure times which occur. Then

$$L = \prod_{fail} \left(F(t_i) - F(t_{i-1})\right)^{d_i} \prod_i \left(1 - F(t_i)\right)^{c_i}$$

So we have showed that amongst all cdf's with fixed values $F(t_i)$ at the failure times t_i, then the discrete cdf has the maximum likelihood, amongst those with d_i failures at t_i and c_i censorings in the interval $[t_i, t_{i+1})$.

Let us consider the **discrete case** and let

$$\mathrm{P}\big\{\text{fail at } t_i | \text{survived to } t_i-\big\} = h_i$$

Then

$$\begin{aligned} S(t_i) &= 1 - F(t_i) = \prod_1^i (1-h_j), \\ f(t_i) &= h_i \prod_1^{i-1} (1-h_j) \end{aligned}$$

Finally we have

$$L = \prod_{t_i} h_i^{d_i} (1-h_i)^{n_i - d_i}$$

where n_i is the number at risk at time t_i. This is usually referred to as the number in the risk set.

Note

$$n_{i+1} + c_i + d_i = n_i$$

11.4.2 Kaplan-Meier estimator

This estimator for $S(t)$ uses the mle estimators for h_i. Taking logs

$$l = \sum_i d_i \log h_i + \sum_i (n_i - d_i) \log(1-h_i)$$

Differentiate with respect to h_i

$$\begin{aligned} \frac{\partial l}{\partial h_i} &= \frac{d_i}{h_i} - \frac{n_i - d_i}{1-h_i} = 0 \\ &\Longrightarrow \quad \widehat{h_i} = \frac{d_i}{n_i} \end{aligned}$$

So the Kaplan-Meier estimator is

$$\widehat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

where

$$\begin{aligned} n_i &= \#\{\text{in risk set at } t_i\}, \\ d_i &= \#\{\text{events at } t_i\}. \end{aligned}$$

Note that $c_i = \#\{\text{censored in } [t_i, t_{i+1})\}$. If there are no censored observations before the first failure time then $n_0 = n_1 = \#\{\text{in study}\}$. Generally we assume $t_0 = 0$.

11.4.3 Nelson-Aalen estimator and new estimator of S

The Nelson-Aalen estimator for the cumulative hazard function is

$$\widehat{H}(t) = \sum_{t_i \le t} \frac{d_i}{n_i} \quad \left(= \sum_{t_i \le t} \widehat{h_i} \right)$$

This is natural for a discrete estimator, as we have simply summed the estimates of the hazards at each time, instead of integrating, to get the cummulative hazard. This correspondingly gives an estimator of S of the form

$$\begin{aligned} \widetilde{S}(t) &= \exp\left(-\widehat{H}(t)\right) \\ &= \exp\left(-\sum_{t_i \le t} \frac{d_i}{n_i}\right) \end{aligned}$$

It is not difficult to show by comparing the functions $1-x, \exp(-x)$ on the interval $0 \le x \le 1$, that $\widetilde{S}(t) \ge \widehat{S}(t)$.

11.4.4 Invented data set

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

$$2, 5, 5, 6+, 7, 7+, 12, 14+, 14+, 14+ \tag{11.1}$$

Here + indicates a censored observation. Then we can calculate both estimators for $S(t)$ at all time points. It is considered unsafe to extrapolate much beyond the last time point, 14, even with a large data set.

Table 11.1: Computations of survival estimates for invented data set (11.1)

t_i	d_i	n_i	$\hat{h}_i$	$\hat{S}(t_i)$	$\widetilde{S}(t_i)$
2	1	10	0.10	0.90	0.90
5	2	9	0.22	0.70	0.72
6	1	6	0.17	0.58	0.63
12	1	4	0.25	0.44	0.54

Lecture 12

Confidence intervals and left truncation

We need to find confidence intervals (pointwise) for the estimators of $S(t)$ at each time point. We differentiate the log-likelihood and use likelihood theory,

$$l = \sum_i d_i \log h_i + \sum_i (n_i - d_i) \log(1 - h_i),$$

differentiated twice to find the Hessian matrix $\left\{ \frac{\partial^2 l}{\partial h_i \partial h_j} \right\}$.

Note that since l is a sum of functions of each individual hazard the Hessian must be diagonal.

The estimators $\left\{ \widehat{h_1}, \widehat{h_2}, \ldots, \widehat{h_n} \right\}$ are asymptotically unbiased and are asymptotically jointly normally distributed with approximate variance I^{-1}, where the information matrix is given by

$$I = \mathbf{E}\left(- \left\{ \frac{\partial^2 l}{\partial h_i \partial h_j} \right\} \right).$$

Since the Hessian is diagonal, the covariances are all asymptotically zero, and coupled with asymptotic normality, this ensures that all pairs $\widehat{h_i}, \widehat{h_j}$ are asymptotically independent.

$$-\frac{\partial^2 l}{\partial h_i^2} = \frac{d_i}{h_i^2} + \frac{n_i - d_i}{(1 - h_i)^2}$$

We use the observed information J and so replace h_i in the above by its estimator $\widehat{h_i} = \frac{d_i}{n_i}$. Hence we have

$$\mathbf{var}\, \widehat{h_i} \approx \frac{d_i \left(n_i - d_i \right)}{n_i^3}.$$

12.1 Greenwood's formula

12.1.1 Reminder of the δ method

If the random variation of Y around μ is small (for example if μ is the mean of Y and $\mathbf{var}Y$ has order $\frac{1}{n}$), we use:

$$g(Y) \approx g(\mu) + (Y - \mu) g'(\mu) + \frac{1}{2} (Y - \mu)^2 g''(\mu) + \ldots$$

Taking expectations

$$\begin{aligned}\mathbf{E}(g(Y)) &= g(\mu) + O\left(\frac{1}{n}\right)\\ \mathbf{var}(\mathbf{g}(\mathbf{Y})) &= \mathbf{g}'(\mu)^2\mathbf{var}Y + o\left(\frac{1}{n}\right)\end{aligned}$$

12.1.2 Derivation of Greenwood's formula for var$(\widehat{S}(t))$

$$\log \widehat{S}(t) = \sum_{t_i \le t} \log\left(1 - \widehat{h_i}\right)$$

But

$$\mathbf{var}\ \widehat{h_i} \approx \frac{d_i\,(n_i - d_i)}{n_i^3} \quad \text{and } \widehat{h_i} \xrightarrow{P} h_i$$

so that, given $g(h_i) = \log(1 - h_i)$,

$$g'(h_i) = \frac{-1}{(1 - h_i)}$$

we have

$$\begin{aligned}\mathbf{var}\ \log\left(1 - \widehat{h_i}\right) &\approx \frac{1}{(1-h_i)^2}\mathbf{var}\ \widehat{h_i}\\ &\approx \frac{1}{(1-\frac{d_i}{n_i})^2}\frac{d_i\,(n_i - d_i)}{n_i^3}\\ &= \frac{d_i}{n_i\,(n_i - d_i)}\end{aligned}$$

Since $\widehat{h_i}, \widehat{h_j}$ are asymptotically independent we can put all this together to get

$$\mathbf{var}\log\left(\widehat{S}(t)\right) = \sum_{t_i \le t} \frac{d_i}{n_i\,(n_i - d_i)} \tag{12.1}$$

Let $Y = \log \widehat{S}$ and note that we need $\mathbf{var}\left(e^Y\right) \approx \left(e^Y\right)^2 \mathbf{var}Y$, again using the delta-method. Finally we have *Greenwood's formula*

$$\mathbf{var}\left(\widehat{S}(t)\right) \approx \widehat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i\,(n_i - d_i)}\ . \tag{12.2}$$

Applying this to the same sort of argument to the Nelson-Aalen estimator and its extension to the survival function we also see

$$\mathbf{var}\ \widehat{H}(t) \approx \sum_{t_i \le t} \frac{d_i\,(n_i - d_i)}{n_i^3}$$

and

$$\begin{aligned}\mathbf{var}\widetilde{S}(t) &= \mathbf{var}\left(\exp(-\widehat{H}(t)\right)\\ &\approx \left(e^{-H}\right)^2 \sum_{t_i \le t} \frac{d_i\,(n_i - d_i)}{n_i^3}\\ &\approx \left(\widetilde{S}(t)\right)^2 \sum_{t_i \le t} \frac{d_i\,(n_i - d_i)}{n_i^3}\end{aligned}$$

Clearly these estimates are only reasonable if each n_i is sufficiently large, since they rely heavily on asymptotic calculations.

12.2 Left truncation

Left truncation is easily dealt with in the context of nonparametric survival estimation. Suppose the invented data set comes from the following hidden process: There is an event time, and an independent censoring time, and, in addition, a truncation time, which is the time when that individual becomes available to be studied. For example, suppose this were a nursing home population, and the time being studied is the number of years after age 80 when the patient first shows signs of dementia. The censoring time might be the time when the person dies or moves away, or when the study ends. The study population consists of those who have entered the nursing home free of dementia. The truncation time would be the age at which the individual moves into the nursing home.

Table 12.1: Invented data illustrating left truncation. Event times after the censoring time may be purely nominal, since they may not have occurred at all; these are marked with *. The row *Observation* shows what has actually been observed. When the event time comes before the truncation time the individual is not included in the study; this is marked by a ∘.

Patient ID	5	2	9	0	1	3	7	6	4	8
Event time	2	5	5	*	7	*	12	*	*	*
Censoring time	10	8	7	8	11	7	14	14	14	14
Truncation time	−2	3	6	0	1	0	6	6	−5	1
Observation	2	5	∘	8+	7	7+	12	14+	14+	14+

Table 12.2: Computations of survival estimates for invented data set of Table 12.1.

t_i	d_i	n_i	$\hat{h}_i$	$\hat{S}(t_i)$	$\widetilde{S}(t_i)$
2	1	6	0.17	0.83	0.85
5	1	6	0.17	0.69	0.72
7	1	7	0.14	0.58	0.62
12	1	4	0.25	0.45	0.48

We give a version of these data in Table 12.1. Note that patient number 9 was truncated at time 6 (i.e., entered the nursing home at age 86) but her event was at time 5 (i.e., she had already suffered from dementia since age 85), hence was not included in the study. In table 12.2 we give the computations for the Kaplan-Meier estimate of the survival function. The computations are exactly the same as those of section 11.4.4, except for one important change: The number at risk n_i is not simply the number $n - \sum_{t_i<t} d_i - \sum_{t_i<t} k_i$ of individuals who have not yet had their event or censoring time. Rather, an individual is at risk at time t if her event time and censoring time are both $\geq t$, and if the truncation time is $\leq t$. (As usual, we assume that individuals who have their event or are censored in a given year, were at risk during that year. We are similarly assuming that those who entered the study at age x are at risk during that year.) At the start of our invented study there are only 6 individuals at risk, so the estimated hazard for the event at age 2 becomes 1/6.

In the most common cases of truncation we need do nothing at all, other than be careful in interpreting the results. For instance, suppose we were simply studying the age after 80 at which individuals develop dementia by a longitudinal design, where 100 healthy individuals 80 years old are recruited and followed for a period of time. Those who are already impaired at age 80 are truncated. All this means is that we have to understand (as we surely would) that the results are conditional on the individual not suffering from dementia until age 80.

We can compute variances for the Kaplan-Meier and Nelson-Aalen estimators using Greenwood's formula exactly as before, only taking care to use the reinterpreted number at risk. The one problem that arises is that individuals may enter into the study slowly, yielding a small number at risk, and hence very wide error bounds, which of course will carry through to the end.

12.3 Example: The AML study

In the 1970s it was known that individuals who had gone into remission after chemotherapy for acute lymphatic leukemia would benefit — by longer remission times — from a course of continuing "maintenance" chemotherapy. A study [EEH+77] pointed out that "Despite a lack of conclusive evidence, it has been assumed that maintenance chemotherapy is useful in the management of acute myelogenous leukemia (AML)." The study set out to test this assumption, comparing the duration of remission between an experimental group that received the additional chemotherapy, and a control group that did not. (This analysis is based on the discussion in [MGM01].)

The data are from a preliminary analysis of the data, before completion of the study. The duration of complete remission in weeks was given for each patient (11 maintained, 12 non-maintained controls); those who were still in remission at the time of the analysis are censored observations. The data are given in Table 12.3. They are included in the `survival` package of `R`, under the name `aml`.

Table 12.3: Times of complete remission for preliminary analysis of AML data, in weeks. Censored observations denoted by +.

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

The first thing we do is to estimate the survival curves. The summary data and computations are given in Table 12.4. The Kaplan-Meier survival curves are shown in Figure 12.1. In Table 12.5 we show the computations for confidence intervals just for the Kaplan-Meier curve of the maintenance group. The confidence intervals are based on the logarithm of survival, using (12.1) directly. That is, the bounds on the confidence interval are

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

where z is the appropriate quantile of the normal distribution. Note that the approximation cannot be assumed to be very good in this case, since the number of individuals at risk is too small for the asymptotics to be reliable. We show the confidence intervals in Figure 12.2.

Table 12.4: Computations for the Kaplan-Meier and Nelson-Aalen survival curve estimates of the AML data.

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

Important: The estimate of the variance is more generally reliable than the assumption of normality, particularly for small numbers of events. Thus, the first line in Table 12.5 indicates that the estimate of $\log \hat{S}(9)$ is associated with a variance of 0.009. The error in this estimate is on the order of n^{-3}, so it's potentially about 10%. On the other hand, the number of events observed has binomial distribution, with parameters around $(11, 0.909)$, so it's very far from a normal distribution. We could improve our confidence interval by using the Poisson confidence intervals worked out in Problem Sheet 3, question 2, or binomial confidence interval. We will not go into the details in this course.

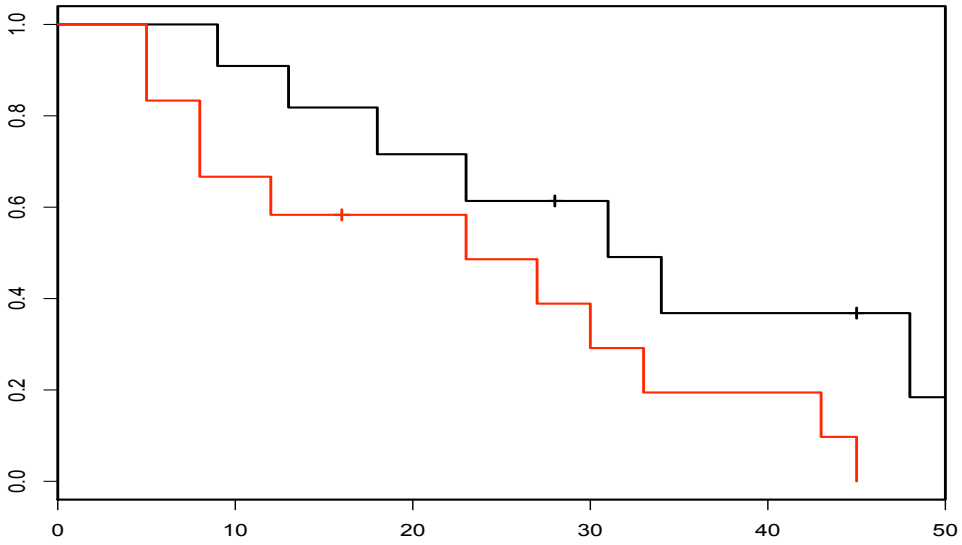


Figure 12.1: Kaplan-Meier estimates of survival in maintenance (black) and non-maintenance groups in the AML study.

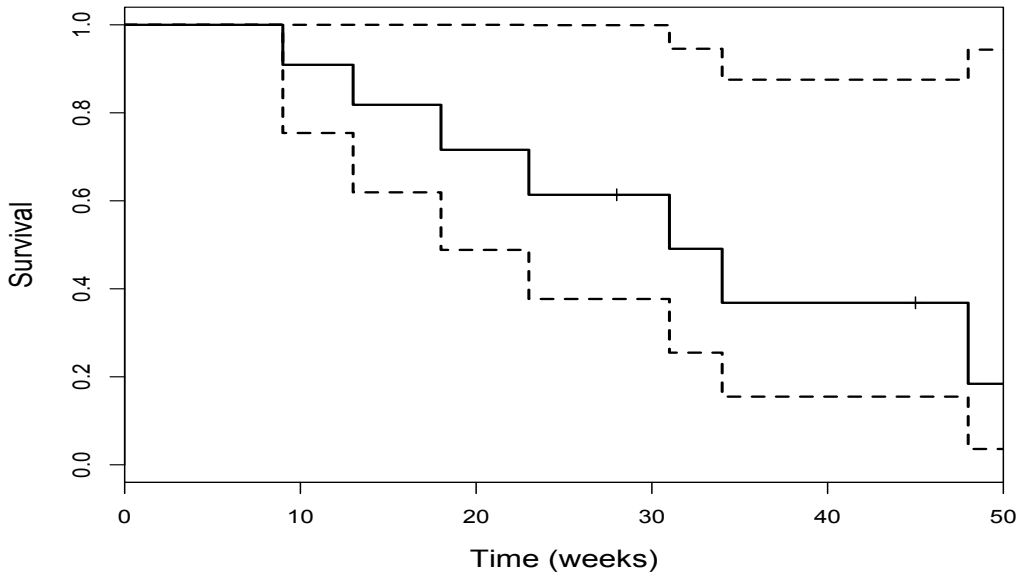


Figure 12.2: Greenwood's estimate of 95% confidence intervals for survival in maintenance group of the AML study.

Table 12.5: Computations for Greenwood's estimate of the standard error of the Kaplan-Meier survival curve from the maintenance population in the AML data. "lower" and "upper" are bounds for 95% confidence intervals, based on the log-normal distribution.

t_i	n_i	d_i	$\frac{d_i}{n_i(n_i-d_i)}$	$\mathrm{Var}(\log \hat{S}(t_i))$	lower	upper
9	11	1	0.009	0.009	0.754	1.000
13	10	1	0.011	0.020	0.619	1.000
18	8	1	0.018	0.038	0.488	1.000
23	7	1	0.024	0.062	0.377	0.999
31	5	1	0.050	0.112	0.255	0.946
34	4	1	0.083	0.195	0.155	0.875
48	2	1	0.500	0.695	0.036	0.944

12.4 Actuarial estimator

The **actuarial estimator** is a further estimator for $S(t)$. It is given as

$$S^*(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i - \frac{1}{2}c_i}\right)$$

The intervals between consecutive failure times are usually of constant length, and it is generally used by actuaries and demographers following a cohort from birth to death. Age will normally be the time variable and hence the unit of time is 1 year.

Lecture 13

Semiparametric models: accelerated life, proportional hazards

Reading: Cox & Oakes chapter 5.1–5.7, K & M chapter 8.1–8.4, 8.8, 12.1–5

13.1 Introduction to semiparametric modeling

We learned in section 6.3 how to compare observed mortality to a standard life table. In many settings, though, we are interested to compare observed mortality (or more general event times) between groups, or between individuals with different values of a quantitative covariate, and in the presence of censoring. For example,

Often we are interested to compare two (or more) different lifetime distributions. An approach that has been found to be effective is to think of there being a "standard" lifetime which may be modified in various simple ways to produce the lifetimes of the subpopulations. The standard lifetime is commonly estimated nonparametrically, while the modifications — usually the characteristic of primary interest — is reduced to one or a few parameters. The modifications may either involve a discrete collection of parameters — one parameter for each of a small number of subpopulations — or a regression-type parameter multiplied by a continuous covariate.

Examples of the former type would be clinical trials, where we compare survival time between treatment and control groups, or an observational study where we compare survival rates of smokers and non-smokers. An example of the second time would be testing time to appearance of full-blown AIDS symptoms as a function of measured T-cell counts.

There are two popular general classes of model as in the heading above - AL and PH.

13.2 Accelerated Life models

Suppose there are (several) groups, labelled by index i. The accelerated life model has a survival curve for each group defined by

$$S_i(t) = S_0(\rho_i t)$$

where $S_0(t)$ is some baseline survival curve and ρ_i is a constant specific to group i.

If we plot S_i against $\log t$, $i = 1, 2, \ldots, k$, then we expect to see a horizontal shift as

$$S_i(t) = S_0(\mathrm{e}^{\log \rho_i + \log t}) \, .$$

13.2.1 Medians and Quantiles

Note too that each group has a different median lifetime, since, if $S_0(m) = 0.5$,

$$S_i(\frac{m}{\rho_i}) = S_0(\rho_i \frac{m}{\rho_i}) = 0.5,$$

giving a median for group i of $\frac{m}{\rho_i}$. Similarly if the $100\alpha\%$ quantile of the baseline survival function is t_α, then the $100\alpha\%$ quantile of group i is $\frac{t_\alpha}{\rho_i}$.

13.3 Proportional Hazards models

In this model we assume that the hazards in the various groups are proportional so that

$$h_i(t) = \rho_i h_0(t)$$

where $h_0(t)$ is the baseline hazard. Hence we see that

$$S_i(t) = S_0(t)^{\rho_i}$$

Taking logs twice we get

$$\log\left(-\log S_i(t)\right) = \log \rho_i + \log\left(-\log S_0(t)\right)$$

So if we plot the RHS of the above equation against either t or $\log t$ we expect to see a vertical shift between groups.

13.3.1 Plots

Taking both models together it is clear that we should plot

$$\log\Big(-\log \widehat{S}_i(t)\Big) \text{ against } \log t$$

as then we can check for *AL and PH in one plot.* Generally $\widehat{S}_i$ will be calculated as the Kaplan-Meier estimator for group i, and the survival function estimator for each group will be plotted on the same graph.

(i) If the accelerated life model is plausible we expect to see a horizontal shift between groups.

(ii) If the proportional hazards model is plausible we expect to see a vertical shift between groups.

13.4 AL parametric models

There are several well-known parametric models which have the accelerated life property. These models also allow us to take account of continuous covariates such as blood pressure.

Name	Survival $S(t)$	Hazard $h(t)$	Density $f(t) = h(t)S(t)$
Weibull	$\exp(-(\rho t)^{\alpha})$	$\alpha\rho^{\alpha}t^{\alpha-1}$	$\alpha\rho^{\alpha}t^{\alpha-1}\mathrm{e}^{-(\rho t)^{\alpha}}$
log-logistic	$\frac{1}{1+(\rho t)^{\alpha}}$	$\frac{\alpha\rho^{\alpha}t^{\alpha-1}}{1+(\rho t)^{\alpha}}$	$\frac{\alpha\rho^{\alpha}t^{\alpha-1}}{(1+(\rho t)^{\alpha})^2}$
log-normal	1-$\Phi\left(\frac{\log t+\log\rho}{\sigma}\right)$	$\cdots$	$\frac{1}{t\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(\log t+\log\rho)^2\right)$
exponential	$e^{-\rho t}$	ρ	$\rho e^{-\rho t}$

Remarks:

(i) Exponential is a submodel of Weibull with $\alpha = 1$

(ii) log-normal is derived from a normal distribution with mean $-\log\rho$ and variance σ^2. In this distribution $\alpha = \frac{1}{\sigma}$ has the same role as in the Weibull and log-logistic.

(iii) The **shape** parameter is α. The **scale** parameter is ρ.

Shape in the hazard function $h(t)$ is important.

Weibull $\cdots$ $\begin{cases} h \text{ monotonic increasing } \alpha>1 \\ h \text{ monotonic decreasing } \alpha<1 \end{cases}$

log-normal $\cdots$ $h \longrightarrow 0$ as $t \longrightarrow 0, \infty$, one mode only

log-logistic $\cdots$ see problem sheet 5.

Comments:

a) to get a "bathtub" shape we might use a mixture of Weibulls. This gives high initial probability of an event, a period of low hazard rate and then increasing hazard rate for larger values of t.

b) to get an inverted "bathtub" shape we may have a mixture of log-logistics, or possibly a single log-normal or single log-logistic.

To check for appropriate parametric model (given AL checked)

There are some distributional ways of testing for say Weibull v. log-logistic etc., but they involve generalised F-distributions and are not in general use.

We can do a simple test for Weibull v. exponential as this simply means testing a null hypothesis $\alpha = 1$, and the exponential is a sub-model of the Weibull model. Hence we can use the likelihood ratio statistic which involves

$$2\log\widehat{L}_{weib} - 2\log\widehat{L}_{\exp} \sim \chi^2(1), \text{ asymptotically.}$$

13.4.1 Plots for parametric models

However most studies use plots which give a rough guide from shape. We should use a **straightline fit** as this is the fit which the human eye spots easily.

1. **Exponential** - $S = \mathrm{e}^{-\rho t}$, plot $\log S$ v. t
2. **Weibull** - $S = \mathrm{e}^{-(\rho t)^{\alpha}}$, plot $\log(-\log S)$ v. $\log t$
3. **log-logistic** - $S = \frac{1}{1+(\rho t)^{\alpha}}$, plot $\cdots$ see problem sheet 6
4. **log-normal** - $S = 1 - \Phi\left(\frac{\log t+\log\rho}{\sigma}\right)$, plot $\Phi^{-1}(1-S)$ v. $\log t$ or equivalently $\Phi^{-1}(S)$ v. $\log t$

In each of the above we would estimate S with the Kaplan-Meier estimator $\widehat{S}(t)$, and use this to construct the plots.

13.4.2 Regression in parametric AL models (assuming right censoring only)

In general studies each observation will have measured explanatory factors such as age, smoking status, blood pressure and so on. We need to incorporate these into a model using some sort of generalised regression. It is usual to do so by making ρ a function of the explanatory variables. For each observation (say individual in a clinical trial) we set the scale parameter $\rho = \rho(\beta \cdot x)$, where $\beta \cdot x$ is a linear predictor composed of a vector x of known explanatory variables (covariates) and an unknown vector β of parameters which will be estimated. The most common link function is

$$\log \rho = \beta \cdot x \ , \ \text{ equivalently } \ \rho = \mathrm{e}^{\beta \cdot x} \ .$$

Censoring is assumed to be independent mechanism and is sometimes referred to as non-informative.

The **shape parameter** α is assumed to be the same for each observation in the study.

There are often very many covariates measured for each subject in a study.

A row of data will have perhaps:-

response - event time t_i , status δ_i (=1 if failure, =0 if censored)

covariates - age, sex, systolic blood pressure, treatment, and so a mixture of categorical variables and continuous variables amongst the covariates.

Suppose that Weibull is a good fit. Then

$$\begin{aligned} S(t) &= \mathrm{e}^{-(\rho t)^\alpha} \ \text{ and } \ \rho = \mathrm{e}^{\beta \cdot x} \\ \beta.x &= b_0 + b_1 x_{age} + b_2 x_{sex} + b_3 x_{sbp} + b_4 x_{trt} \end{aligned}$$

where b_0 is the intercept and all regression coefficients b_i are to be estimated, as well as estimating α. Note this model assumes that α is the same for each subject. We have not shown, but could have, interaction terms such as $x_{age} * x_{trt}$. This interaction would allow a different effect of age according to treatment group.

Suppose subject j has covariate vector x_j and so scale parameter

$$\rho_j = \mathrm{e}^{\beta \cdot x_j} \ .$$

This gives a likelihood

$$\begin{aligned} L(\alpha, \beta) &= \prod_j \left(\alpha \rho_j^\alpha t_j^{\alpha - 1} \right)^{\delta_j} \mathrm{e}^{-(\rho_j t_j)^\alpha} \\ &= \prod_j \left(\alpha \mathrm{e}^{\alpha \beta \cdot x_j} t_j^{\alpha - 1} \right)^{\delta_j} \mathrm{e}^{-\left(\mathrm{e}^{\beta \cdot x_j} t_j \right)^\alpha} . \end{aligned}$$

We can now compute MLEs for α and all components of the vector β, using numerical optimisation, giving estimators $\widehat{\alpha}$, $\widehat{\beta}$ together with their standard errors ($=\sqrt{\mathrm{var}\widehat{\alpha}}, \sqrt{\mathrm{var}\widehat{\beta}_j}$) calculated from the observed information matrix (see problem sheet 5). Of course, the same could have been done for another parametric model instead of the Weibull.

As already noted we can test for $\alpha = 1$ using

$$2 \log \widehat{L}_{weib} - 2 \log \widehat{L}_{\exp} \sim \chi^2(1), \text{ asymptotically.}$$

Packages allow for Weibull, log-logistic and log-normal models, sometimes others.

13.4.3 Linear regression in parametric AL models

The idea is to mirror ordinary linear regression and find a baseline distribution which does not depend on ρ, similar to looking at the error term in least squares regression. We give the derivation just for the Weibull distribution, but similar arguments work for all AL parametric models. We have

$$\begin{aligned} S(t) &= \mathrm{e}^{-(\rho t)^\alpha} = \mathrm{P}\big\{T > t\big\} \\ &= \mathrm{P}\big\{\log T > \log t\big\} \\ &= \mathrm{P}\big\{\alpha\left(\log T + \log\rho\right) > \alpha\left(\log t + \log\rho\right)\big\} \end{aligned}$$

Now let $Y = \alpha\left(\log T + \log\rho\right)$ and $y = \alpha\left(\log t + \log\rho\right)$.

$$\begin{aligned} \mathrm{P}\big\{Y &> y\big\} = S_Y(y) \\ &= S(t) \\ &= \mathrm{e}^{-(\rho t)^\alpha} \\ &= \exp(-\mathrm{e}^y) \end{aligned}$$

Hence we have

$$\log T = -\log\rho + \frac{1}{\alpha}Y, \quad \text{where} \quad S_Y(y) = \exp(-\mathrm{e}^y)$$

The distribution of Y is independent of the parameters ρ and α. And in the case of the Weibull distribution its distribution is called the **extreme value distribution** and is as above.

In general we will write $\log T = -\log\rho + \frac{1}{\alpha}Y$ for all AL parametric models, and Y has a distribution in each case which is independent of the model parameters.

Name	$S(t)$	Y	$S_Y(y)$ distribution
Weibull	$\exp(-(\rho t)^\alpha)$	$\log T = -\log\rho + \frac{1}{\alpha}Y$	$\exp(-\mathrm{e}^y)$: extreme value distrib.
log-logistic	$\frac{1}{1+(\rho t)^\alpha}$	$\log T = -\log\rho + \frac{1}{\alpha}Y$	$(1+\mathrm{e}^y)^{-1}$: logistic distribution
log-normal	1-$\Phi\left(\frac{\log t+\log\rho}{\sigma}\right)$	$\log T = -\log\rho + \sigma Y$	$1-\Phi(y)$: $\mathcal{N}(0,1)$

as before $\alpha = \frac{1}{\sigma}$, for the log-normal.

In recent years, a semi-parametric model has been developed in which the baseline survival function S_0 is modelled non-parametrically, and each subject has time t scaled to $\rho_j t$. This model is beyond the scope of this course.

Lecture 14

Cox regression, Part I

Again each subject j has a vector of covariates x_j and scale parameter $\rho_j = \rho_j(\beta \cdot x_j)$. The basic assumption is that any two subjects have hazard functions whose ratio is a constant proportion which depends on the covariates. Hence we may write

$$h_j(t) = \rho_j h_0(t)$$

where h_0 is the baseline hazard function, β is a vector of regression coefficients to be estimated, and ρ_j again depends on the linear predictor $\beta.x_j$.

A general link could be used but in **Cox regression** $\rho_j = \mathrm{e}^{\beta.x_j}$. This model is termed semi-parametric because the functional form of the baseline hazard is not given, but is determined from the data, similarly to the idea for estimating the survival function by the Kaplan-Meier estimator.

14.1 What is Cox Regression?

Cox regression is Proportional Hazards with a semi-parametric model.

Suppose the event times are given by $0 < t_1 < t_2 < \cdots < t_m$. At this stage we assume no tied event times (list does not include censored times).

Let $[i]$ denote the subject with event at t_i.

Definition: Risk Set

The risk set R_i is the set of those subjects available for the event at time t_i.

Reminder: if we know that there are d subjects with hazard functions $h_1, \cdots, h_d$ then, knowing there is an event at time t_0, the probability that subject j has the event is

$$\mathrm{P}\big\{\text{subject } \ j \,\big|\, t_0\big\} = \frac{h_j(t_0)}{h_1(t_0) + \cdots + h_d(t_0)} \ .$$

Under the proportional hazards assumption we have

$$\mathrm{P}\big\{[\mathrm{i}] \,\big|\, t_i\big\} = \frac{\rho_{[i]} h_0(t_i)}{\sum\limits_{j \in R_i} \rho_j h_0(t_i)} = \frac{\rho_{[i]}}{\sum\limits_{j \in R_i} \rho_j}$$

and the probability that $[i]$ has the event given it occurs at time t_i no longer depends on t_i.

Under the Cox regression model we have

$$\mathrm{P}\big\{[\mathrm{i}] \,\big|\, t_i\big\} = \frac{\mathrm{e}^{\beta \cdot x_{[i]}}}{\sum\limits_{j \in R_i} \mathrm{e}^{\beta \cdot x_j}} \ .$$

This probability only depends on the order in which subjects have the events.

The idea of the model is to specify a partial likelihood which depends only on the order in which events occur, not the times at which they occur. This means that the functional form of h_0, the baseline hazard function, is not required.

Definition: Partial Likelihood

$$L_P(\beta) = \prod_{t_i} \frac{e^{\beta \cdot x_{[i]}}}{\sum_{j \in R_i} e^{\beta \cdot x_j}}$$

where R_i is the risk set at t_i, and subject $[i]$ is the subject with the event at t_i.

We can think of the partial likelihood as the *joint density function for subjects' ranks in terms of event order*, if there were no censoring and no tied event times.

Consequently if we use the partial likelihood for estimation of parameters we are *losing information,* because we are suppressing the actual times of events even though they are known, hence the name "partial likelihood".

Interestingly the partial likelihood acts in an exactly similar manner to the likelihood. Compute $\widehat{\beta}_P$ such that

$$L_P\left(\widehat{\beta}_P\right) = \sup_{\beta} \prod_{t_i} \frac{e^{\beta \cdot x_{[i]}}}{\sum_{j \in R_i} e^{\beta \cdot x_j}}$$

Then $\widehat{\beta}_P$ maximises the partial likelihood and has all the usual properties.

Properties

(i) $\widehat{\beta}_P \xrightarrow{P} \beta$ as $m \longrightarrow \infty$ (and hence the number in the study tends to infinity also),
(ii) $\mathbf{var}\widehat{\beta}_P \approx I_P^{-1}$, where I_P is calculated from L_P in exactly the same way as for the usual information and likelihood,
(iii) asymptotic normality of $\widehat{\beta}_P$ also holds.

There are journal papers showing that the % information lost by ignoring actual event times is smaller than one might expect. All of the above rests on the assumption that the Cox regression model fits the data, of course.

14.2 Relative Risk

There is a big difference between deductions from AL parametric analysis and PH semi-parametric analysis. In PH the intercept is non-identifiable and so we are estimating relative risk between subjects, not absolute risk, when we estimate the model parameters.

Definition: relative risk

The relative risk at time t between two subjects with covariates x_1, x_2 and hazard functions h_2, h_1 is defined to be

$$\frac{h_2(t)}{h_1(t)} .$$

For the Cox regression model this becomes time independent and is given by

$$e^{\beta.(x_2 - x_1)} .$$

The intercept is non-identifiable because

$$h(t;x) = e^{\beta \cdot x} h_0(t) = e^{\alpha + \beta \cdot x} \left(e^{-\alpha} h_0(t)\right)$$

for any α. This means that any such intercept α included with the regression expression $\beta \cdot x$ simply cancels out in the partial likelihood. Hence an intercept is never included in the linear regressor in this model.

14.3 Baseline hazard

However we do need to estimate the cumulative baseline hazard function and also the baseline survival function.

Definition: Breslow's estimator for the baseline cumulative hazard function
Suppose the baseline survival is given by

$$\widehat{S_0}(t) = \mathrm{e}^{-\widehat{H_0}(t)},$$

where the discrete hazard estimation $\widehat{h_0}$ is given by

$$\widehat{h_0}(t_i) = \frac{1}{\sum\limits_{j \in R_i} \mathrm{e}^{\beta \cdot x_j}}$$

Breslow's estimator is given by

$$\widehat{h_0} = \frac{1}{\sum\limits_{j \in R_i} \mathrm{e}^{\beta \cdot x_j}} \tag{14.1}$$

In some sense the discrete estimates for $\widehat{h_0}(t_i)$ can be thought of as the maximum likelihood estimators from the full likelihood, provided we assume that the hazard distribution is discrete (which of course it generally is not). When $\hat{\beta} = 0$ or when the covariates are all 0, this reduces simply to the Nelson-Aalen estimator. Otherwise, we see that this is equivalent to a modified Nelson-Aalen estimator, where the size of the risk set is weighted by the relative risks of the individuals. In other words, the estimate of $\widehat{h_0}$ is equivalent to the standard estimate # events/time at risk, but now time at risk is weighted by the relative risk.

The estimator may be loosely derived as follows:

$$\begin{aligned}
\ell(h) &= \sum_{t_i} \log(1 - e^{-h_{[i]}(t_i)}) - \sum_{\substack{j \in R_i \\ j \neq [i]}} h_j \\
&= \sum_{t_i} \log(1 - e^{-\hat{\rho}_{[i]} h_0(t_i)}) - \sum_{\substack{j \in R_i \\ j \neq [i]}} \hat{\rho}_j h_0(t_i).
\end{aligned}$$

We estimate $h_0(t_i)$ by

$$\begin{aligned}
0 &= \frac{\hat{\rho}_{[i]} e^{-\hat{\rho}_{[i]} \hat{h}_0(t_i)}}{1 - e^{-\hat{\rho}_{[i]} \hat{h}_0(t_i)}} - \sum_{\substack{j \in R_i \\ j \neq [i]}} \hat{\rho}_j \\
&\approx \frac{\hat{\rho}_{[i]}(1 - \hat{\rho}_{[i]} \hat{h}_0(t_i))}{\hat{\rho}_{[i]} \hat{h}_0(t_i)} - \sum_{\substack{j \in R_i \\ j \neq [i]}} \hat{\rho}_j \\
&= \frac{(1 - \hat{\rho}_{[i]} \hat{h}_0(t_i))}{\hat{h}_0(t_i)} - \sum_{\substack{j \in R_i \\ j \neq [i]}} \hat{\rho}_j.
\end{aligned}$$

(In the second line we have assumed $h_0(t_i)$ to be small.) Thus

$$1 \approx \hat{h}_0(t_i) \left(\sum_{j \in R_i} \hat{\rho}_j \right),$$

which is the same as (14.1).

Lecture 15

Cox regression, Part II

15.1 Dealing with ties

Until now in this section we have been assuming that the times of events are all distinct. In situations where event times are equal, we can carry out the same computations for Cox regression, only using a modified version of the partial likelihood. Suppose R_i is the set of individuals at risk at time t_i, and D_i the set of individuals who have their event at that time. We assume that the ties are not real ties, but only the result of discreteness in the observation. Then the probability of having precisely those individuals at time t_i will depend on the order in which they actually occurred. For example, suppose there are 5 individuals at risk at the start, and two of them have their events at time t_1. If the relative risks were $\{\rho_1, \ldots, \rho_5\}$, where $\rho_j = e^{\beta \cdot x_j}$, then the first term in the partial likelihood would be

$$\frac{\rho_1}{\rho_1+\rho_2+\rho_3+\rho_4+\rho_5} \cdot \frac{\rho_2}{\rho_2+\rho_3+\rho_4+\rho_5} + \frac{\rho_2}{\rho_1+\rho_2+\rho_3+\rho_4+\rho_5} \cdot \frac{\rho_1}{\rho_1+\rho_3+\rho_4+\rho_5}.$$

The number of terms is $d_i!$, so it is easy to see that this computation quickly becomes intractable.

A very good alternative — accurate and easy to compute — was proposed by B. Efron. Observe that the terms differ in the denominator merely by a small change due to the individuals lost from the risk set. If the deaths at time t_i are not a large proportion of the risk set, then we can approximate this by deducting the average of the risks that depart. In other words, in the above example, the first contribution to the partial likelihood becomes

$$\frac{\rho_1\rho_2}{(\rho_1+\rho_2+\rho_3+\rho_4+\rho_5)(\frac{1}{2}(\rho_1+\rho_2)+\rho_3+\rho_4+\rho_5)}.$$

More generally, the partial likelihood becomes

$$L_P(\beta) = \prod_{t_i} e^{\beta \cdot \sum_{j \in D_i} x_j} \prod_{k=0}^{d_i-1} \left(\sum_{j \in R_i} e^{\beta \cdot x_j} - \frac{k}{d_i} \sum_{j \in D_i} e^{\beta \cdot x_j} \right)^{-1}.$$

We take the same approach to estimating the baseline hazard:

$$\hat{h}_0(t_i) = \sum_{k=0}^{d_i-1} \left(\sum_{j \in R_i} e^{\hat{\beta} \cdot x_j} - \frac{k}{d_i} \sum_{j \in D_i} e^{\hat{\beta} \cdot x_j} \right)^{-1}.$$

Another approach, due to Breslow, makes no correction for the progressive loss of risk in the denominator:

$$L_P^{Breslow}(\beta) = \prod_{t_i} e^{\beta \cdot \sum_{j \in D_i} x_i} \left(\sum_{j \in R_i} e^{\beta \cdot x_i} \right)^{-d_i}.$$

This approximation is always too small, and tends to shift the estimates of β toward 0. It is widely used as a default in software packages (`SAS`, not `R`!) for purely historical reasons.

15.2 Plot for PH assumption with continuous covariate

Suppose we have a continuous covariate and we wish to check the proportional hazards assumption for that covariate. We do not have natural groups of subjects with the same value of that covariate.

Provided there is sufficient data we would group the subjects in quintiles of the covariate. Then we have 5 groups and can find the Kaplan-Meier estimator for each group. As before we plot

$$\log(-\log(\widehat{S_k}(t))) \text{ v. } \log t$$

for each $k = 1, \cdots, 5$ on the same graph. There should be a roughly constant vertical separation of groups. It generally is not a wonderful method, but is better than nothing.

15.3 The AML example

We continue looking at the leukemia study that we started to consider in section 12.3. First, in Figure 15.1 we plot the iterated logarithm of survival against time, to test the proportional hazards assumption. The PH assumption corresponds to the two curves differing by a vertical shift. The result makes this assumption at least credible.

We code the data with covariate $x = 0$ for the maintained group, and $x = 1$ for the non-maintained group. Thus, the baseline hazard will correspond to the maintained group, and e^β will be the relative risk of the non-maintained group. From Table 12.4 we see that the Efron approximate partial likelihood is given by

$$\begin{aligned} L_P(\beta) = &\left(\frac{e^{2\beta}}{(12e^\beta+11)(11e^\beta+11)}\right)\left(\frac{e^{2\beta}}{(10e^\beta+11)(9e^\beta+11)}\right) \\ &\times \left(\frac{1}{8e^\beta+11}\right)\left(\frac{e^\beta}{8e^\beta+10}\right)\left(\frac{1}{7e^\beta+10}\right)\left(\frac{1}{6e^\beta+8}\right) \\ &\times \left(\frac{e^\beta \cdot 1}{(6e^\beta+7)(5.5e^\beta+6.5)}\right)\left(\frac{e^\beta}{5e^\beta+6}\right)\left(\frac{e^\beta}{4e^\beta+5}\right) \\ &\times \left(\frac{1}{3e^\beta+5}\right)\left(\frac{e^\beta}{3e^\beta+4}\right)\left(\frac{1}{2e^\beta+4}\right)\left(\frac{e^\beta}{2e^\beta+3}\right)\left(\frac{e^\beta}{e^\beta+3}\right)\left(\frac{1}{2}\right) \end{aligned} \tag{15.1}$$

A plot of $L_P(\beta)$ is shown in Figure 15.4.

In the one-dimensional setting it is straightforward to estimate $\hat{\beta}$ by direct computation. We see the maximum at $\hat{\beta} = 0.9155$ in the plot of Figure 15.4. In more complicated settings, there are good maximisation algorithms built in to the `coxph` function in the `survival` package of `R`. Applying this to the current problem, we obtain:

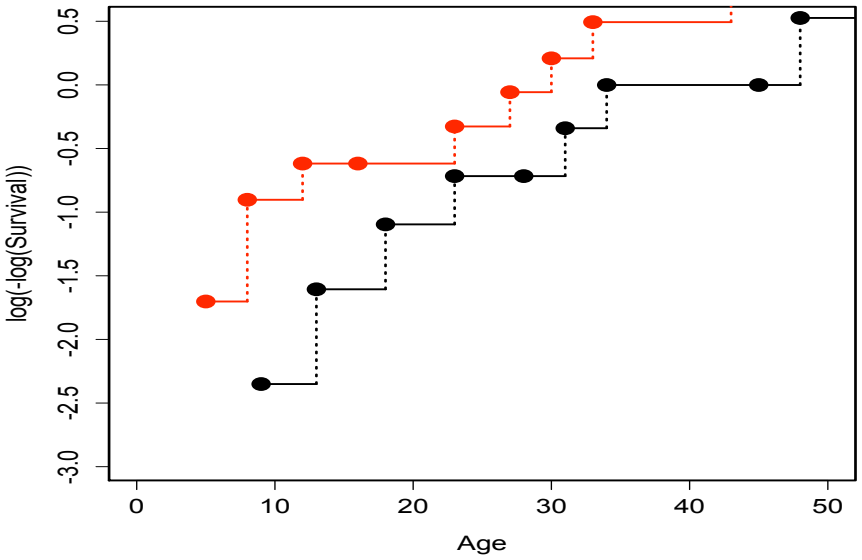


Figure 15.1: Iterated log plot of survival of two populations in AML study, to test proportional hazards assumption.

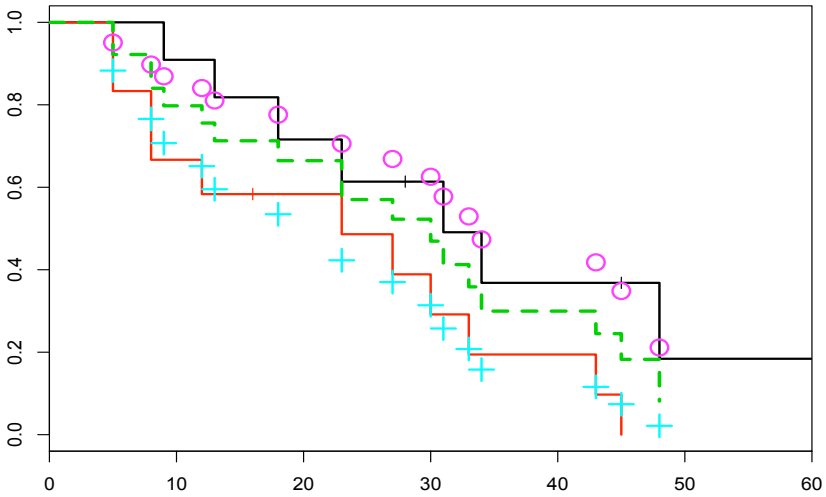


Figure 15.2: Estimated baseline hazard under the PH assumption. The purple circles show the baseline hazard; blue crosses show the baseline hazard shifted up proportionally by a multiple of $e^{\hat{\beta}} = 2.5$. The dashed green line shows the estimated survival rate for the mixed population (mixing the two estimates by their proportions in the initial population).

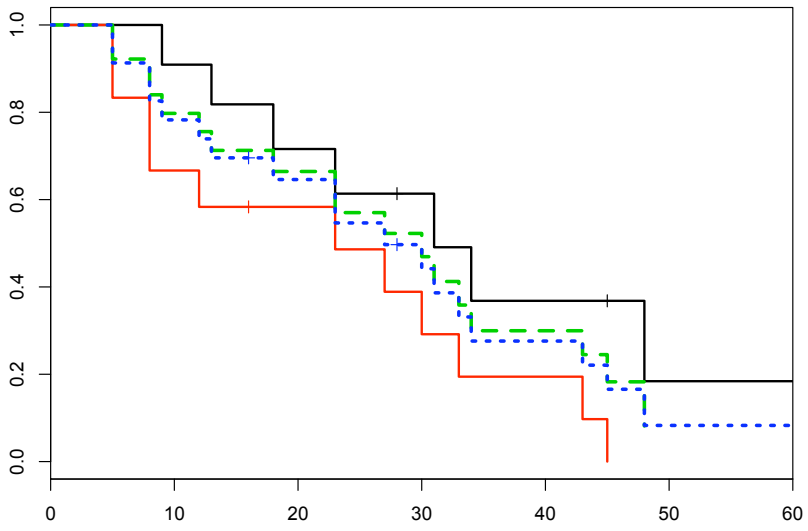


Figure 15.3: Comparing the estimated population survival under the PH assumption (green dashed line) with the estimated survival for the combined population (blue dashed line), found by applying the Nelson-Aalen estimator to the population, ignoring the covariate.

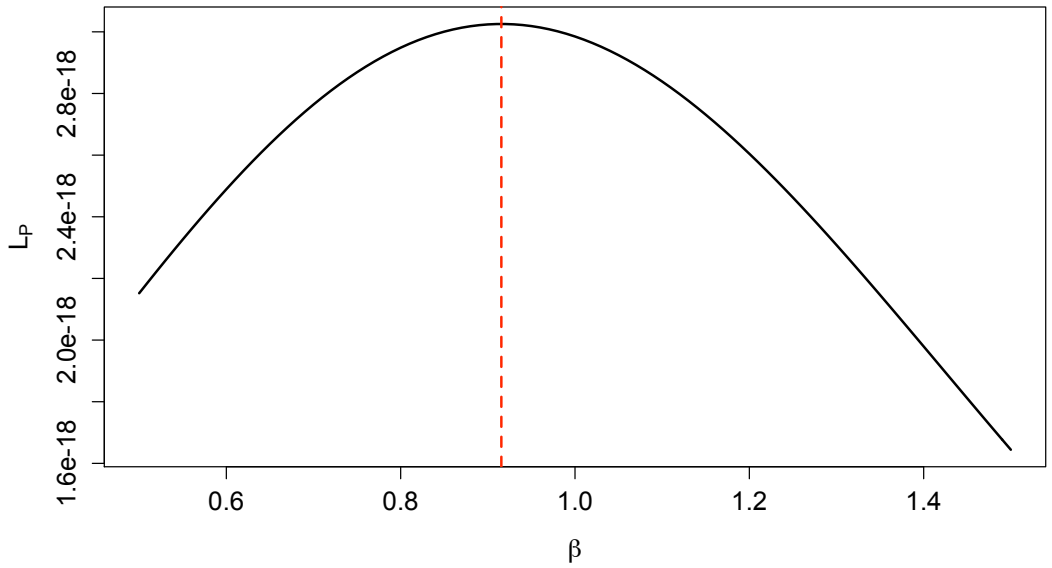


Figure 15.4: A plot of the partial likelihood from (15.1). Dashed line is at $\beta = 0.9155$.

Table 15.1: Output of the `coxph` function run on the `aml` data set.

coxph(formula = Surv(time, status) ∼ x, data = aml)					
	coef	exp(coef)	se(coef)	z	p
×Nonmaintained	0.916	2.5	0.512	1.79	0.074
Likelihood ratio test=3.38 on 1 df			p=0.0658	n= 23	

The z is simply the Z-statistic for testing the hypothesis that $\beta = 0$, so $z = \hat{\beta}/SE(\hat{\beta})$. We see that $z = 1.79$ corresponds to a p-value of 0.074, so we would not reject the null hypothesis at level 0.05.

We show the estimated baseline hazard in Figure 15.2; the relevant numbers are given in Table 15.2. For example, the first hazard, corresponding to $t_1 = 5$, is given by

$$\hat{h}_0(5) = \frac{1}{12e^{\hat{\beta}} + 11} + \frac{1}{11e^{\hat{\beta}} + 11} = 0.050,$$

substituting in $\hat{\beta} = 0.9155$.

Table 15.2: Computations for the baseline hazard LME for the AML data, in the proportional hazards model, with maintained group as baseline, and relative risk $e^{\hat{\beta}} = 2.498$.

	Maintenance		Non-Maintenance (control)		Baseline		
t_i	n_i^M	d_i^M	n_i^N	d_i^N	$\hat{h}_0(t_i)$	$\hat{H}_0(t_i)$	$\widetilde{S}_0(t_i)$
5	11	0	12	2	0.050	0.050	0.951
8	11	0	10	2	0.058	0.108	0.898
9	11	1	8	0	0.032	0.140	0.869
12	10	0	8	1	0.033	0.174	0.841
13	10	1	7	0	0.036	0.210	0.811
18	8	1	6	0	0.043	0.254	0.776
23	7	1	6	1	0.095	0.348	0.706
27	6	0	5	1	0.054	0.403	0.669
30	5	0	4	1	0.067	0.469	0.625
31	5	1	3	0	0.080	0.549	0.577
33	4	0	3	1	0.087	0.636	0.529
34	4	1	2	0	0.111	0.747	0.474
43	3	0	2	1	0.125	0.872	0.418
45	3	0	1	1	0.182	1.054	0.348
48	2	1	0	0	0.500	1.554	0.211

Lecture 16

Testing Hypotheses

Reading: C& O sections 8.6–8.7, K & M sections 7.1–7.3

A common question that we may have is, whether two (or more) samples of survival times may be considered to have been drawn from the same distribution: That is, whether the populations under observation are subject to the same hazard rate.

16.1 Tests in the regression setting

1) A package will produce a test of whether or not a regression coefficient is 0. It uses properties of mle's. Let the coefficient of interest be b say. Then the null hypothesis is $\mathrm{H_O} : b = 0$ and the alternative is $\mathrm{H}_A : b \neq 0$. At the 5% significance level, $\mathrm{H_O}$ will be accepted if the *p-value* $p > 0.05$, and rejected otherwise.

2) In an AL parametric model if α is the shape parameter then we can test $\mathrm{H_O} : \log \alpha = 0$ against the alternative $\mathrm{H}_A : \log \alpha \neq 0$. Again mle properties are used and a p-value is produced as above. In the case of the Weibull if we accept $\log \alpha = 0$ then we have the simpler exponential distribution (with $\alpha = 1$).

3) We have already mentioned that, to test Weibull v. exponential with null hypothesis $\mathrm{H_O}$: exponential is an acceptable fit, we can use

$$2 \log \widehat{L}_{weib} - 2 \log \widehat{L}_{\exp} \sim \chi^2(1), \text{ asymptotically.}$$

16.2 Non-parametric testing of survival between groups

16.2.1 General principles

We will consider only the case where the data splits into two groups. There is a relatively easy extension to $k > 2$ groups.

We define the following notation

$$\begin{aligned} &\text{Event times are } 0 < t_1 < t_2 < \cdots < t_m. \\ \text{For } i = 1, 2, \ldots, m, \text{ and } j = 1, 2, &\, d_{ij} = \#\text{ events at } t_i \text{ in group } j, \\ &\, n_{ij} = \#\text{ in risk set at } t_i \text{ from group } j, \\ &\, d_i = \#\text{ events at } t_i, \\ &\, n_i = \#\text{ in risk set at } t_i. \end{aligned}$$

Thus, when the number of groups $k = 2$, we have $d_i = d_{i1} + d_{i2}$ and $n_i = n_{i1} + n_{i2}$.

Generally we are interested in testing the null hypothesis H_0, that there is no difference between the hazard rates of the two groups, against the two-sided alternative that there **is** a difference in the hazard rates. The guiding principle is quite elementary, quite similar to our approach to the proportional hazards model: We treat each event time t_i as a new and independent experiment. Under the null hypothesis, the next event is simply a random sample from the risk set. Thus, the probability of the death at time t_i being from group 1 is n_{i1}/n_i, and the probability of it being from group 2 is n_{i2}/n_i.

This describes only the setting where the events all occur at distinct times: That is, d_i are all exactly 1. More generally, the null hypothesis predicts that the group identities of the individuals whose events are at time t_i are like a sample of size d_i without replacement from a collection of n_{i1} '1's and n_{i2} '2's. The distribution of d_{i1} under such sampling is called the hypergeometric distribution. It has

$$\text{expectation} = d_i \frac{n_{i1}}{n_i}, \text{ and}$$
$$\text{variance} =: \sigma_i^2 = \frac{n_{i1} n_{i2} (n_i - d_i) d_i}{n_i^2 (n_i - 1)}.$$

Note that if d_i is negligible with respect to n_i, this variance formula reduces to $d_i(\frac{n_{i1}}{n_i})(\frac{n_{i2}}{n_i})$, which is just the variance of a binomial distribution.

Conditioned on all the events up to time t_i (hence on n_i, n_{i1}, n_{i2}) and on d_i, the random variable $d_{i1} - n_{i1}\frac{d_i}{n_i}$ has expectation 0 and variance σ_i^2. If we multiply it by an arbitrary weight $W(t_i)$, determined by the data up to time t_i, we still have $W(t_i)(d_{i1} - n_{i1}\frac{d_i}{n_i})$ being a random variable with (conditional) expectation 0, but now (conditional) variance $W(t_i)^2\sigma_i^2$. This means that if we define for $k = 1, \ldots, m$

$$M_k := \left(\sum_{i=1}^{k} W(t_i)\big(d_{i1} - n_{i1}\frac{d_i}{n_i}\big) \right)_{k=1}^{m},$$

these will be random variables with expectation 0 and variance $\sum_{i=1}^{k} W(t_i)^2\sigma_i^2$. While the increments are not independent, we may still apply a version of the Central Limit Theorem to show that M_k is approximately normal when the sample size is large enough. (In technical terms, the sequence of random variables M_k is a *martingale*, and the appropriate theorem is the Martingale Central Limit Theorem. See [HH80] for more details.) We then base our tests on the statistic

$$Z := \frac{\sum_{i=1}^{m} W(t_i)\big(d_{i1} - n_{i1}\frac{d_i}{n_i}\big)}{\sqrt{\sum_{i=1}^{m} W(t_i)^2 \frac{n_{i1} n_{i2} (n_i - d_i) d_i}{n_i^2 (n_i - 1)}}},$$

which should have a standard normal distribution under the null hypothesis.

Note that, as in the Cox regression setting, right censoring and left truncation are automatically taken care of, by appropriate choice of the risk sets.

16.2.2 Standard tests

Any choice of weights $W(t_i)$ defines a valid test. Why do we need weights? Since any choice of weights produces a *correct* test, there is no canonical choice. Changing the weights changes the power with respect to different alternatives. Which alternative you choose — hence, which weights you choose — should depend on what deviations from equality you are most interested

in detecting. As always, the test should be chosen beforehand. Multiple testing makes the interpretation of test results problematic.

Some common choices are:

1. $W(t_i) = 1, \forall i$. This is the **log rank test**, and is the test in most common use. The log rank test is aimed at detecting a consistent difference between hazards in the two groups and is best placed to consider this alternative when the proportional hazard assumption applies. It is maximally asymptotically efficient in the proportional hazards context; in fact, it is equivalent to the score test for the Cox regression parameter being 0, hence is asymptotically equivalent to the likelihood ratio test. A criticism is that it can give too much weight to the later event times when numbers in the risk sets may be relatively small.

2. R. Peto and J. Peto [PP72] proposed a test which emphasises deviations that occur early on, when there are more individuals under observation. **Petos' test** uses a weight dependent on a modified estimated survival function, estimated for the whole study. The modified estimator is
$$\widetilde{S}(t) = \prod_{t_i \leq t} \frac{n_i + 1 - d_i}{n_i + 1}$$
and the suggested weight is then
$$W(t_i) = \widetilde{S}(t_{i-1}) \frac{n_i}{n_i + 1}$$
This has the advantage of giving more weight to the early events and less to the later ones where the population remaining is smaller.

3. $W(t_i) = n_i$ has also been suggested (Gehan, Breslow). This again downgrades the effct of the later times.

4. D. Harrington and T. Fleming [HF82] proposed a class of tests that include Petos' test and the logrank test as special cases. The **Fleming-Harrington tests** use
$$W(t_i) = \left(\widehat{S}(t_{i-1})\right)^p \left(1 - \widehat{S}(t_{i-1})\right)^q$$
where $\widehat{S}$ is the Kaplan-Meier survival function, estimated for all the data. Then $p = q = 0$ gives the logrank test and $p = 1, q = 0$ gives a test very close to Peto's test and is called the Fleming-Harrington test. If we were to set $p = 0, q > 0$ this would emphasise the later event times if needed for some reason.

All of these tests may be written in the form
$$\frac{\sum (O_{i1} - E_{i1}) W_i}{\sqrt{\sum \sigma_{i1}^2 W_i^2}},$$
where O_i and E_i are observed and expected numbers of events. Consequently, positive and negative fluctuations can cancel each other out. This could conceal a substantial difference between hazard rates which is not of the proportional hazards form, but where the hazard rates (for instance) cross over, with group 1 having (say) the higher hazard early, and the lower hazard

later. One way to detect such an effect is with a test statistic to which fluctuations contribute only their absolute values. For instance, we could use the standard χ^2 statistic

$$X := \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Asymptotically, this should have the χ^2 distribution with $(k-1)m$ degrees of freedom. Of course, if the number of groups $k = 2$, this is the same as

$$X := \sum_{i=1}^{m} \frac{(O_{i1} - E_{i1})^2}{d_i \frac{n_{i1}}{n_i}(1 - \frac{n_{i1}}{n_i})}.$$

16.3 The AML example

We can use these tests to compare the survival of the two groups in the AML experiment discussed in section 12.3. The relevant quantities are tabulated in Table 16.1.

Time	n_{i1}	n_{i2}	d_{i1}	d_{i2}	σ_i^2	Peto weight
5	11	12	0	2	0.476	0.958
8	11	10	0	2	0.474	0.875
9	11	8	1	0	0.244	0.792
12	10	8	0	1	0.247	0.750
13	10	7	1	0	0.242	0.708
18	8	6	1	0	0.245	0.661
23	7	6	1	1	0.456	0.614
27	6	5	0	1	0.248	0.519
30	5	4	0	1	0.247	0.467
31	5	3	1	0	0.234	0.416
33	4	3	0	1	0.245	0.364
34	4	2	1	0	0.222	0.312
43	3	2	0	1	0.240	0.260
45	3	1	0	1	0.188	0.208

Table 16.1: Data for testing equality of survival in AML experiment.

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

Applying the χ^2 test yields $X = 16.86$, which needs to be compared to χ^2 with 14 degrees of freedom. The resulting p-value is 0.24, which is not at all significant. This should not be seen as surprising: The differences between the two survival curves are clearly mostly in the same direction, so we lose power when applying a test that ignores the direction of the difference.

Bibliography

[Buf77] George Leclerc Buffon. *Essai d'arithmétique morale*. 1777.

[CT406] *CT4: Models Core Reading*. Faculty & Institute of Acutaries, 2006.

[ECIW06] Gregory M. Erickson, Philip J. Currie, Brian D. Inouye, and Alice A. Winn. Tyrannosaur life tables: An example of nonavian dinosaur population biology. *Science*, 313:213–7, 2006.

[EEH+77] Stephen H. Embury, Laurence Elias, Philip H. Heller, Charles E. Hood, Peter L. Greenberg, and Stanley L. Schrier. Remission maintenance therapy in acute myelogenous leukemia. *The Western Journal of Medicine*, 126:267–72, April 1977.

[Fox97] A. J. Fox. English life tables no. 15. Office of National Statistics, London, 1997.

[Gom25] Benjamin Gompertz. On the nature of the function expressive of the law of human mortality and on a new mode of determining life contingencies. *Philosophical transactions of the Royal Society of London*, 115:513–85, 1825.

[HF82] David P. Harrington and Thomas R. Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–66, December 1982.

[HH80] Peter Hall and Christopher C. Heyde. *Martingale Limit Theory and its Application*. Academic Press, New York, London, 1980.

[Kie01] Kathleen Kiernan. The rise of cohabitation and childbearing outside marriage in western Europe. *International Journal of Law, Policy and the Family*, 15:1–21, 2001.

[KT81] Samuel Karlin and Howard M. Taylor. *A Second Course in Stochastic Processes*. Academic Press, 1981.

[Mac96] A. S. Macdonald. An actuarial survey of statistics models for decrement and transition data. I: Multiple state, Poisson and binomial models. *British Actuarial Journal*, 2(1):129–55, 1996.

[ME05] Kyriakos S. Markides and Karl Eschbach. Aging, migration, and mortality: Current status of research on the hispanic paradox. *Journals of Gerontology: Series B*, 60B:68–75, 2005.

[MGM01] Rupert G. Miller, Gail Gong, and Alvaro Muñoz. *Survival Analysis*. Wiley, 2001.

[PP72] Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135(2):185–207, 1972.

[Wac] Kenneth W. Wachter. Essential demographic methods. Unpublished manuscript.