

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

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

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

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

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, [?] 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 different forms of census data. The most common one is

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

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

$$E_x^c = \int_0^n P_{x,t} dt$$

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

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

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

Now assume that, in fact, you are not given d_x but only calendar 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

$$E_x^{c'} = \int_0^n P'_{x,t} dt,$$

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 $k-1 \leq t < k$, but then in $P'_{x+1,k}$ at the next census date. Therefore

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

The ratio $d'_x/E_x^{c'}$ 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 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.

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

The count d_x described in section 5.4 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 $E_x^c := \int_0^T P_{x,t} dt$, where $P_{x,t}$ is the number of individuals alive at time t whose curtate age is x . We can represent this as

$$E_x^c = \int_0^T P_{x,t} dt = \sum_{k=0}^{T-1} \mathbb{P}_{x,k},$$

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

¹These 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/>.)

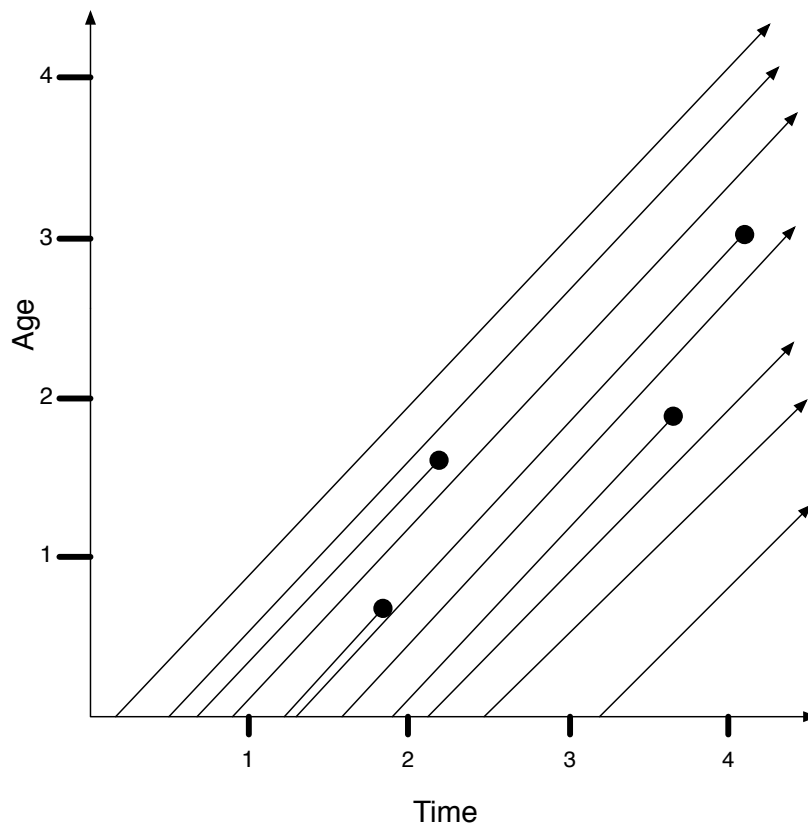


Figure 5.1: A Lexis diagram.

$\frac{1}{2}(P(x, k) + P(x, k + 1))$. Then we get the approximation

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

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.

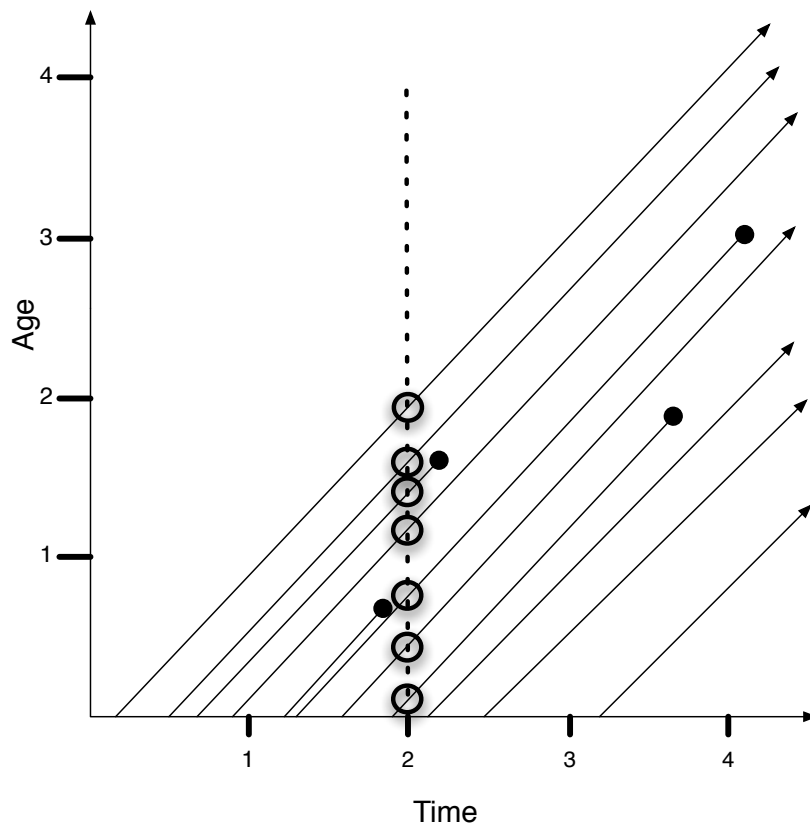


Figure 5.2: Census at time 2 represented by open circles. The population consists of 7 individuals. 4 are between ages 1 and 2, and 3 are between 0 and 1.

- 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

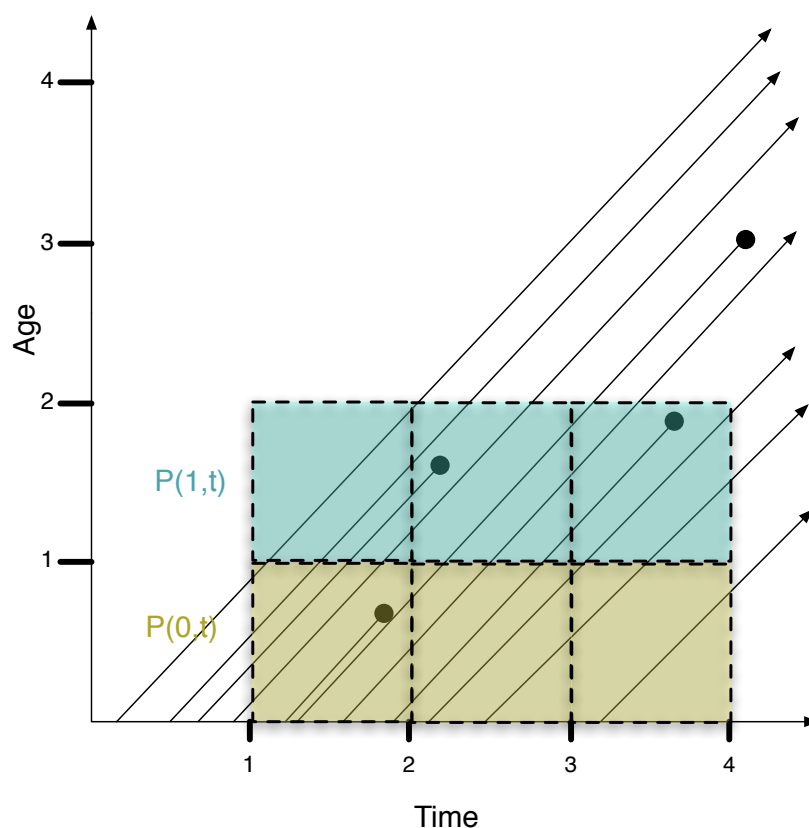


Figure 5.3: Census approximation when events are counted by actual curtate age. The vertical segments represent census counts.

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.

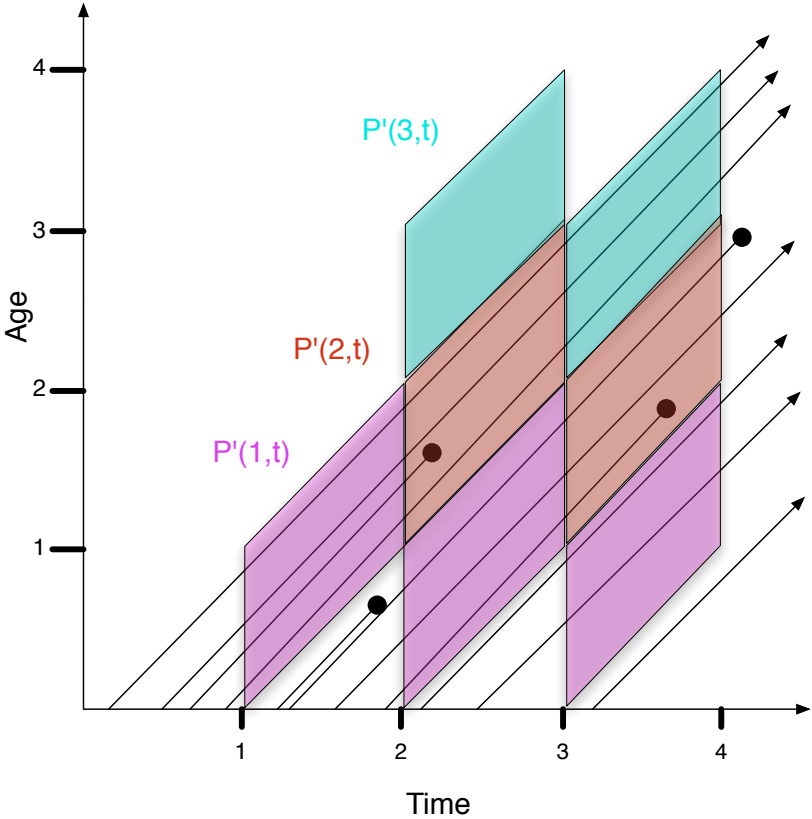


Figure 5.4: Census approximation when events are counted by calendar year of birth and death. Vertical segments bounding the coloured regions represent census counts.