"Lexis cohorts": Extracting information on half-year cohorts from 1-year format Lexis data

Alan Cohen and John Tillinghast, Johns Hopkins University

Short abstract

Often, demographic studies across countries may require cohorts at a finer scale than the 1-year cohorts available in sources like the Human Mortality Database. Examples include effects of season of birth on life expectancy or the discrete cohort effects of events such as pandemics and wars. We have developed a method to test differences between half-year cohorts when data are available at a 1-year time scale in Lexis format. For every birth year and death age, there are two possible death years. Those dying in the earlier year have a ~75% chance of being born in the first half of the birth year; the converse is true for those dying in the later year. We consider those dying in the earlier and later year separate "Lexis" cohorts for statistical purposes and test differences between them. Effect sizes cannot generally be estimated, but qualitative differences can be detected.

Long Abstract

Traditional demographic analyses rely on making a clear choice between using period and cohort analyses. If we have data on death year and age at death, we have period data; if we have data on birth year and age we have cohort data. The difference arises because a person born in year t (say, 1920) who dies at age a (say, 60) could die in either year $a+t$ or $a+t+1$ (1980 or 1981), depending on when the birthday fell in the birth year and how close the person made it to age $a+1$. For example, someone dying at age 60.95 is still recorded as dying at age 60, but probably died in 1981. When we wish to assess the effect of being born in a given year, we use cohort data; when we wish to assess the effect of events during a year on deaths across age-classes, we use period data.

The problem that can arise here is one of scale. We might wish sometimes to detect a pattern with a clear threshold point that may be partway through a year, or with cyclical effects operating more rapidly than can be detected by dividing the data up by year. If we have only period and cohort data by year, there is generally no way to get information on a finer scale. However, sometimes we have Lexis data available, in which case both birth year and death year are present. This provides some additional information that we should be able to utilize for detection of finer scale patterns; the topic of this article is how to maximize the information we can gain by utilizing Lexis data.

At first glance, it is not immediately apparent how we can get more information. For example, our 1920-born 60-year-old can now be definitively assigned to, say, 1981 as a death year, but we cannot say with certainty when in 1920 she was born or when in 1980 she died. She could have been born early in 1920 (if she died even earlier in 1981) or have died late in 1981 (if she was born even later in 1920). So, unlike our source data, we cannot use absolute knowledge of birth or death at finer scale than a year; at best, we can get probabilistic information.

However, the probabilistic information should be relatively accurate, especially if we make two assumptions: (1) that there are roughly equal numbers of births on each day of a year, and (2) that age-specific mortality rates are constant throughout an age-class (e.g., chance of dying at 60.0 is the same as at 60.9). Of course, neither assumption is strictly true, but they are both approximately true, and as we will see their violation should not undermine the validity of our analysis under most circumstances.

The basic insight that allows us to get probabilistic information is that a person born on Jan. 1 of a year is much more likely to die in the first than the second possible death year. If our example subject were born Jan. 1, the only way she could die in 1981 would be to die on Jan. 1 earlier in the day than the time she was born. Given that we only know her age of death to the year, we can calculate this probability. If we know that she was born on Jan. 1, 1920 at 11:59 pm and died at age 60, there is a 1/365 chance that she died in 1981 and a 364/365 chance that she died in 1980 (again, assuming constant death rates from age 60.00 to age 60.99). If she were born on Jan. 2 at 11:59 pm, these probabilities would shift to 2/365 and 363/365 respectively, and so on. The probabilities always sum to 1, of course, and they are linear functions of birth date within a year (Fig. 1).

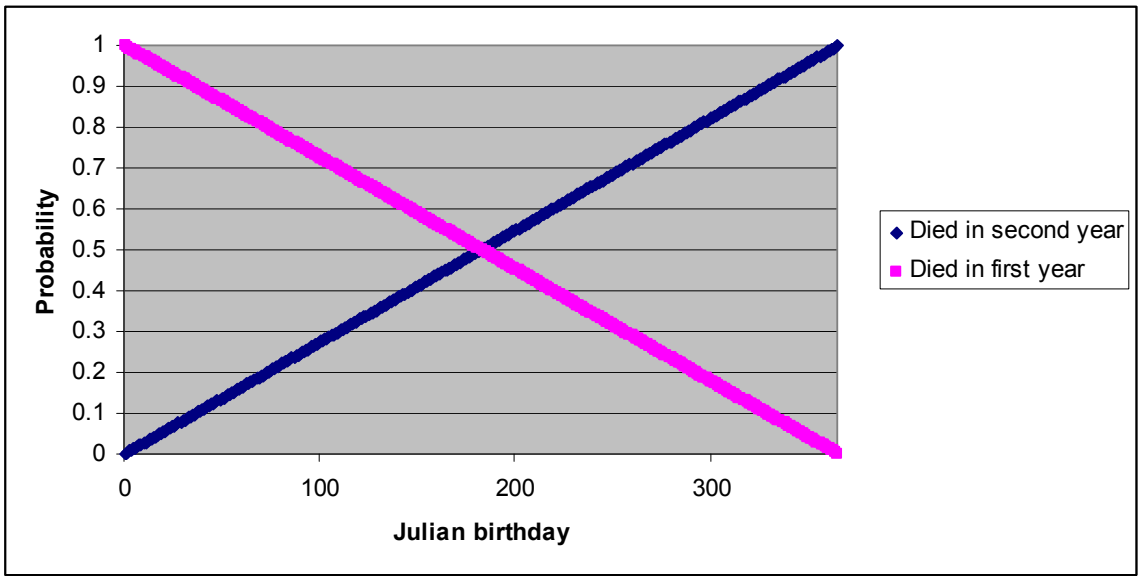

Figure 1

All we know from our Lexis data is that someone died in the earlier or later of the two possible years. However, even with just this knowledge, we can establish the probability distributions of date of birth for these two groups as the lines in Fig. 1. As can be seen, these distributions overlap but are markedly different. The mean date of birth for those who died in the first year is approximately April 1 (1/3 of the way through the year) and for those who died in the second year is Sept. 1 (2/3 of the way through the year). We thus can effectively create two "cohorts" with known mean birthdays. If we conduct analyses on these cohorts, we can now see patterns that emerge at finer scales than one year. Our precision will of course not be as good as if we had two distinct cohorts born with certainty in the first and second half of the year, but if the effects we are looking for are strong enough we should still be able to detect them

The problem of exposures

A major stumbling block to this approach is that we can estimate number of deaths well this way, but we cannot estimate cohort-specific exposures (population size). The data available for exposures are again generally at a scale of one year, but are not available in Lexis format – they are simply number of people of a given age and birth cohort alive in a given year. Thus, in order to calculate a death rate, we need to make an assumption about the distribution of exposures between the two cohorts in a year. The simplest assumption is that the total exposures for a year are equally divided between the two cohorts. This implies that there is no substantial seasonality in birth rate. Although birth rate is known to change seasonally, the two cohorts may have roughly equal representation across summer and winter (with the former containing more spring births and the latter containing more fall births). One could easily incorporate known patterns in seasonality of birth to weight for this.

A larger issue would be a cohort-specific event that affects cohort size. For example, wars or other dramatic events can affect birth rates, and if this effect is stronger in one half of the year, the exposures could be significantly different for the two cohorts, and this effect could easily be large enough to bias rate estimates substantially. There is no way to adjust for this unless we have specific information about birth rates at different points in the year. If we do have information on birth rates, we can again easily weight our distribution of exposures between the two cohorts to reflect this.

Relaxing the assumptions

One assumption of the probability distributions for our two cohorts is that there is no seasonality to birth rates. However, seasonality of births will not affect the probability of each year of death for a given birthday; it affects only the distribution of the birthdays. This is easily incorporated into a model. Fig. 2 uses a cosine function to illustrate the sort of effect seasonality could have on our probability distributions. A less symmetric seasonality could bias the original estimates somewhat (Fig. 3). In all cases, seasonality could affect estimates of mean birthday for the two cohorts, but should not be strong enough to change the fundamental difference between an earlier and a later cohort that are approximately 50% distinct.

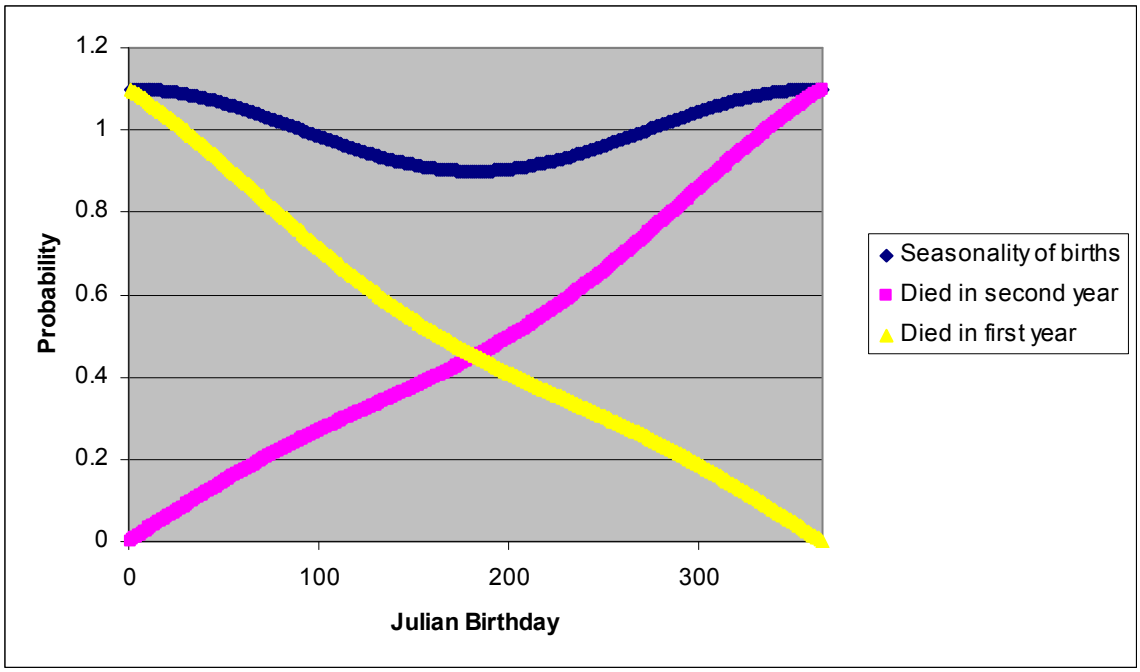

Figure 2

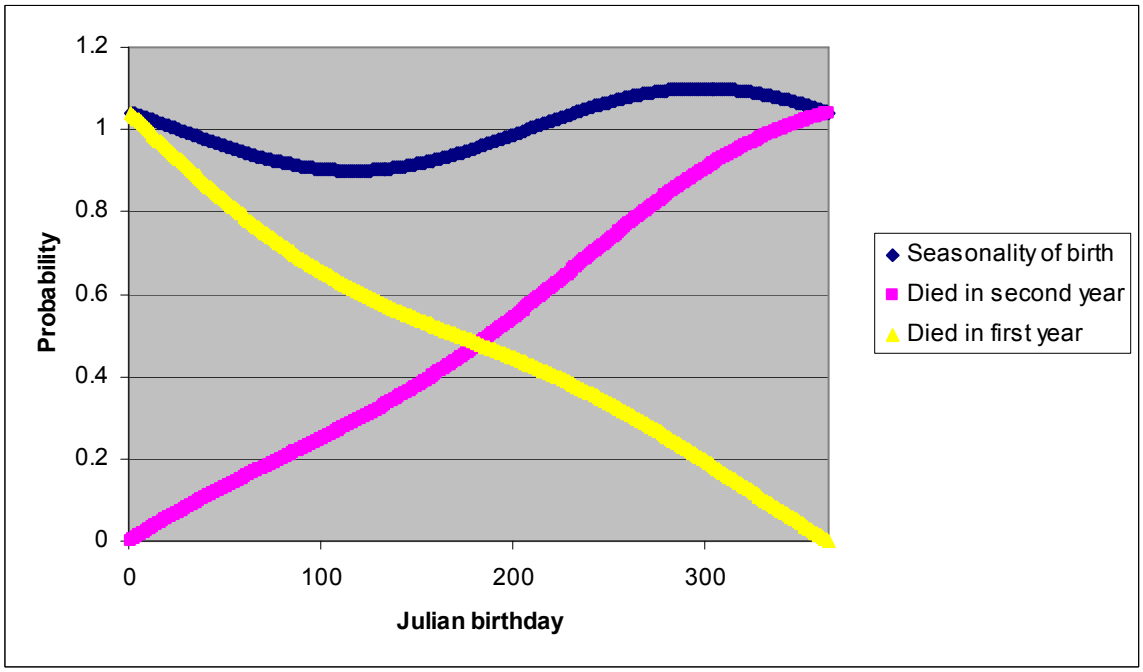

Figure 3

If we allow death rates to vary within an age class, there should be relatively little effect on our model except for infants. Infants have much higher death rates in the first few months of their first year than in the last few months, and this approach should not be used on infant data. However, at older ages death rates change little from year to year. For 60-year olds in Spain from 1916-1922, average death rate is 0.0108, and for 61-year-

olds is 0.0113, a difference of about 5%. This should result in a systematic overestimation of birthday (since people are more likely to have died at an older age and thus to have been born earlier). But again, this overestimation should be less than 5% at most, and the cohorts remain distinct enough for inferences to be made about events happening within a year.

Applications

We have applied this technique to analyses of the effects of developmental exposure to the 1918-1919 flu pandemic on late-life morality and to test the effect of season of birth on late-life mortality. In the former case, the specific timing of the flu made analysis difficult with 1-year cohorts; the negative result with 1-year cohorts was not robust without the finer-scale analysis. In the latter case, effects detected in previous studies by Gabriele Doblhammer and colleagues could be extended to many more countries and time periods.