3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11611

Data granularity in mid-year life table construction

Jose M. Pavía[1], Natalia Salazar[2], Josep Lledó[3]
[1]Universitat de Valencia, Valencia, Spain, [2]Universidad Carlos III, Madrid, Spain, [3]Universidad de Alcalá, Alcalá de Henares, Spain.

Abstract

Life tables have a substantial influence on both public pension systems and life insurance policies. National statistical agencies construct life tables from death rate estimates (m_x), or death probabilities (q_x), after applying various hypotheses to the aggregated figures of demographic events (deaths, migrations and births). The use of big data has become extensive across many disciplines, including population statistics. We take advantage of this fact to create new (more unrestricted) mortality estimators within the family of period-based estimators, in particular, when the exposed-to-risk population is computed through mid-year population estimates. We use actual data of the Spanish population to explore, by exploiting the detailed microdata of births, deaths and migrations (in total, more than 186 million demographic events), the effects that different assumptions have on calculating death probabilities. We also analyse their impact on a sample of insurance product. Our results reveal the need to include granular data, including the exact birthdate of each person, when computing period mid-year life tables.

Keywords: *Mortality tables; mid-year estimators; death rates; big microdata; exposed-to-risk population.*

1. Introduction

Demographic processes affect how populations evolve over time. Understanding these processes is vital for a proper management of social systems, such as pensions and insurance schemes. Exploring mortality dynamics by exploiting all the data at hand is a key factor for improving death probability estimates. Information on variables such as death and birth rates or migratory flows is used by statistical bureaux to construct life tables, which are the cornerstone of the life insurance business.

Since the creation of the first life tables (Graunt, 1662), death probabilities (q_x) and death rates (m_x) have been estimated comparing deaths and exposed-to-risk. For death rates, the numerator refers to the number of deaths with age x in year t, D_x^t, while the denominator accounts for either the total number of 'person-years' at risk (e.g., Wilmoth et al. 2007; Arias, 2015; INE, 2016) or the average population at risk of dying, known as mid-year population estimates (e.g., ONS, 2012). In the era of the IT revolution and the boom of big data, the approach for computing the total number of 'person-years' at risk has been extensively studied in Lledó et al. (2019), while the approach for estimating mid-year figures remains under-researched.

Before the current overabundance of demographic microdata (Ruggles, 2014), period mid-year estimates required assumptions to be computed, such as a uniform distribution of births and deaths and, in some cases, the consideration of a closed population. Nowadays, these assumptions are becoming unnecessary. This paper (i) develops a period mid-year estimator unbounded by a set of hypotheses, (ii) studies (with a real dataset) the impact of using such an estimator on life table construction, and (iii) analyses its effect when calculating premiums for various insurance products.

The rest of the paper is organized as follows. Section 2 describes the methodology and presents the notation. In this section, five different mid-year estimators are discussed. Section 3 briefly introduces the dataset as well as the software used. Section 4 presents the main results, comparing the hypothesis-free estimator we introduce in section 2 with the classical mid-year estimator. In this section, we also analyse the impact of the hypothesis on an insurance product. Section 5 concludes and states future research questions.

2. Methodology

2.1. Preliminary Information

The methodology used counts the demographic variables (births, deaths, migrations, and emigrations) as expressed on the Lexis diagram (Lexi, 1880). Life tables are built based on period estimators, meaning they follow the population over a certain period of time, usually an even number of years. For instance, for construction of British life tables, demographic

events over 3 consecutive years are considered. In this paper, the count is based on the elements of the diagonal portion of the Lexis diagram corresponding to the areas ABFE and DEIH in Figure 1 for one-year tables and on the areas ABFE and DEIH for year 1, GHLK and JKON for year 2 and MNRQ and PQUT for three-year based tables. In both cases, depending on the type of estimator, different assumptions are made to determine which elements are relevant within the area of interest. The specific hypotheses and procedures are further explained in each section.

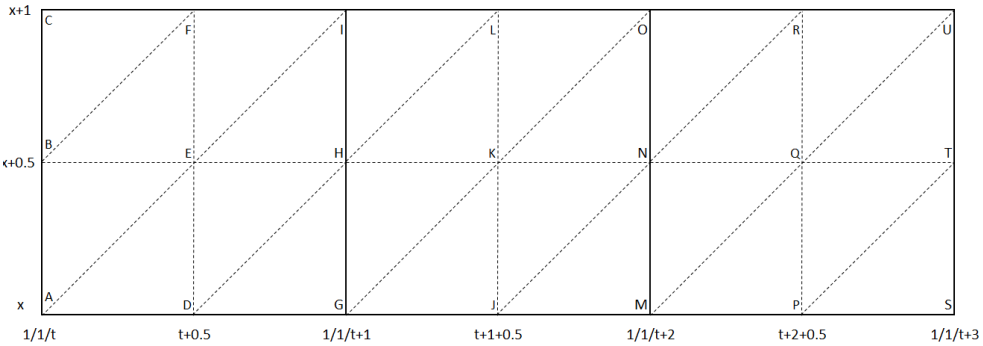

Figure 1. Lexis diagram.

Focusing on the one-year table, mortality rates link the amount of exposed population and deceased individuals through the formula: $\acute{m}_x = D_x^t / C_x^{t+0.5}$, where D_x^t refers to the number of deaths with completed age x in year t and $C_x^{t+0.5}$ to the population with completed age x in the middle of the year t.

We now consider the kind of (micro)data usually available in current statistical systems to develop different estimators by progressively relaxing the hypothesis for their construction until an (apparently) free-hypothesis estimator is reached.

2.2. Closed population and uniform distribution of deaths and birthdates (CP_UD_UB)

In the first scenario, we assume that deaths and birthdates are distributed in a uniform way across every year, implying that each of the triangles that form the Lexis diagram contains the same number of events of interest (1/8 of the total count within the year). Hence, for one period table:

$$\acute{m}_x = \frac{D_x^t}{C_x^{t+\frac{1}{2}}} = \frac{D_x^t}{\frac{1}{2}C_x^t - \frac{2}{8}D_x^t + \frac{1}{2}C_x^{t+1} + \frac{2}{8}D_x^t} = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1}} \qquad (1)$$

and generalizing for *i* periods:

$$\acute{m}_{x,i} = \sum_{j=1}^{i} \left(\frac{D_x^{t+j-1}}{\frac{1}{2} C_x^{t+j-1} + \frac{1}{2} C_x^{t+j}} \right) \qquad (2)$$

To count the people who are alive with completed age x in the middle of year t, represented by segment DF, we start with those alive at the beginning of year t $(t+1)$, which corresponds to segment AB (HI). This segment equals half of segment AC (GI) due to the assumption of uniformity of birthdates, and corresponds to half of the people alive at age t with completed age $(x+1)$, denoted by C_x^t (C_x^{t+1}). The deaths occurring in triangles ABE (DEH) and BFE (EIH) are subtracted (added) to AB (HI) which results in segment EF (DE).

2.3. Open population and uniform distribution of deaths, migrants and births (OP_UD_UM_UB)

In this scenario, the distribution of deaths, birthdates and migratory flows are considered uniform across the year. In this case, the estimator results in the following extension of the basic mortality rate formula:

$$m_x = \frac{D_x^t}{C_x^{t+\frac{1}{2}}} = \frac{D_x^t}{\frac{1}{2} C_x^t - \frac{2}{8} D_x^t + \frac{2}{8} N_x^t + \frac{1}{2} C_x^{t+1} + \frac{2}{8} D_x^t - \frac{2}{8} N_x^t} = \frac{D_x^t}{\frac{1}{2} C_x^t + \frac{1}{2} C_x^{t+1}} \qquad (3)$$

To count the number of people who are alive in the middle of the year, represented by segment DF, we start with those people alive at the beginning of year t $(t+1)$ corresponding to segment AB (GI) which, due to the assumption of the uniformity of births, is equal to half of the people with completed age $(x+1)$, C_x^t (C_x^{t+1}). By taking the difference between the total immigrations and emigrations occurring within the triangles ABE and BFE (DEH and EIH), the net migration flow, noted by N_x^t (N_x^{t+1}) is obtained which, as a result of the uniform distribution of both emigration and immigration, corresponds to 2/8ths of the net migration flow occurring across the whole year. The deaths along with the net migration occurring in the triangles ABE and BFE (DEH and EIH) are subtracted (added) to the count given by AB (HI), which results in segment EF (DE).

Since the net effects of migration for the first and second half of the year cancel out, the simplified version of the equation for this estimator is equal to the final formula of the previous one. The generalization for *i* periods stays the same as before: expression (2).

2.4. Closed population with no hypothesis about the distribution of deaths and uniform distribution of births (CP_NUD_UB)

As the hypothesis of uniformity for the distribution of deaths is relaxed, the same idea of section 2.2 is repeated, still taking segments AB and HI as half of the population with completed age x and $x+1$, respectively, but unlike estimator (1) the number of deaths occurring within areas ABFE and DEIH is not derived from the uniformity assumption. Instead, the exact amount of deceased people in each triangle ABE, BFE, DEH, and EIH is counted and then subtracted or added accordingly. The estimator results in the following extension of the basic mortality rate formula. For one period:

$$\breve{m}_x = \frac{D_x^t}{C_x^{t+\frac{1}{2}}} = \frac{D_x^t}{\frac{1}{2}C_x^t - D_{x,L}^t + \frac{1}{2}C_x^{t+1} + D_{x,R}^t} \qquad (4)$$

generalizing for *i* periods:

$$\breve{m}_{x,i} = \sum_{j=1}^{i}\left(\frac{D_x^{t+j-1}}{\frac{1}{2}C_x^{t+j-1} - D_{x,L}^{t+j-1} + \frac{1}{2}C_x^{t+j} + D_{x,R}^{t+j-1}}\right) \qquad (5)$$

where $D_{x,L}^t$ ($D_{x,R}^t$) is the number of deaths of people with age x during the first (second) half of year t of individuals born in the second (first) half of year $t-x-1$ ($t-x$). This corresponds to area ABFE (DEIH).

2.5. Open population with no hypothesis about the distribution of deaths and migration and uniform distribution of births (OP_NUD_NUM_UB)

When considering the possibility of migration flows, the same procedure of section 2.3 is repeated, still taking segments AB and HI as half of the population with completed age x and $x+1$, respectively. In this case, unlike estimator (4) the number of deaths along with the total net migration flow occurring within areas ABFE and DEIH is not derived using the uniformity assumption. Instead, the exact amount of deceased people and net migrants in each triangle ABE, BFE, DEH, and EIH are counted and then subtracted or added accordingly. Hence, the estimator is written as follows, for one period:

$$\ddot{m}_x = \frac{D_x^t}{C_x^{t+\frac{1}{2}}} = \frac{D_x^t}{\frac{1}{2}C_x^t - D_{x,L}^t - E_{x,L}^t + I_{x,L}^t + \frac{1}{2}C_x^{t+1} + D_{x,R}^t + E_{x,R}^t - I_{x,R}^t} \qquad (6)$$

generalizing for *i* periods:

$$\ddot{m}_{x,i} = \sum_{j=1}^{i}\left(\frac{D_x^{t+j-1}}{\frac{1}{2}C_x^{t+j-1} - D_{x,L}^{t+j-1} - E_{x,L}^{t+j-1} + I_{x,L}^{t+j-1} + \frac{1}{2}C_x^{t+j} + D_{x,R}^{t+j-1} + E_{x,R}^{t+j-1} - I_{x,R}^{t+j-1}}\right) \quad (7)$$

where $E_{x,L}^t$ and $E_{x,R}^t$ ($I_{x,L}^t$ and $I_{x,R}^t$) are the number of emigrations (immigrants) of people with age x during the first and second half of year t of individuals born during the second half of year $t-x-1$ and the first half of $t-x$, respectively. These are areas ABFE and DEIH.

2.6. Open population with no hypothesis about the distribution of deaths, migration, and births (OP_NUD_NUM_NUB)

Finally, when the hypothesis of the uniform distribution of birthdates is not assumed, the same procedure of section 2.5 is repeated, the only difference being that unlike estimator (6) the count for segments AB and HI is not estimated by means of the uniform hypothesis. Instead, the exact amount of people alive in AB and HI is counted using a summation function. This last estimator is synthesized as follows, for one period:

$$\ddot{m}_x = \frac{D_x^t}{C_x^{t+\frac{1}{2}}} = \frac{D_x^t}{\sum_{d=0}^{0.5} C_{x,d}^t - D_{x,L}^t - E_{x,L}^t + I_{x,L}^t + \sum_{d=0.5}^{1} C_{x,d}^t + D_{x,R}^t + E_{x,R}^t - I_{x,R}^t} \quad (8)$$

generalizing for *i* periods:

$$\ddot{m}_{x,i} = \sum_{j=1}^{i}\left(\frac{D_x^{t+j-1}}{\sum_{d=0}^{0.5} C_{x,d}^{t+j-1} - D_{x,L}^{t+j-1} - E_{x,L}^{t+j-1} + I_{x,L}^{t+j-1} + \sum_{d=0.5}^{1} C_{x,d}^{t+j-1} + D_{x,R}^{t+j-1} + E_{x,R}^{t+j-1} - I_{x,R}^{t+j-1}}\right) \quad (9)$$

where $\sum_{d=0}^{0.5} C_{x,d}^{t+j-1}$ is the exact number of people with completed age x $(x+1)$ alive at the beginning of year t $(t+1)$, corresponding to the segment AB (HI).

3. Data and Software

The microdata used to make the computations was purchased from the Spanish National Institute of Statistics (INE). The database consists of the number of deaths, emigrations, immigrations and birthdates; the day, month and year of each event are specified as well as the gender of the individual. This information is available for the period 2005-2008. Overall, we processed, analysed and dealt with more than 180.15 million population inputs, 1.5 million death inputs, 0.7 million of emigrant inputs and 3.2 million of immigrant inputs,

covering both genders. In total, more than 186 million demographic events were studied individually. All the estimators were calculated using the statistical software R, version 3.6.1 (R Core Team, 2019).

4. Results

Throughout the previous section, we have gradually incorporated more detailed information into the mid-year mortality estimator until a hypothesis-free estimator is reached. In this section, using data from 2005 to 2007 to resemble the British mid-year estimator, we compare the most restricted three-periods version of the estimator (CP_UD_UB) with the three-periods hypothesis-free estimator (OP_NUD_NUM_NUB). The formulas for these two estimators correspond to equations (2) and (9), respectively, when $i = 3$.

As the accuracy of the estimator increases, so does the computation time. To check if it is worth exploiting the detailed microdata, we proceed to investigate issue (ii). To do so, we use the absolute relative discrepancy, $m_x - \ddot{m}_x \vee \quad / \ddot{m}_x$, as dissimilarity statistic and compare the life tables calculated using estimators (2) and (9) with $i = 3$. Figure 2 shows the differences.

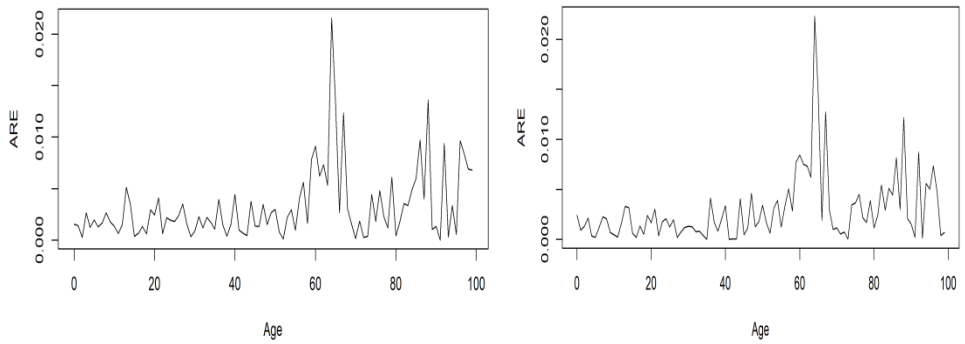

Figure 2. Absolute relative discrepancies between death rates computed using the CP_UD_UB and OP_NUD_NUM_NUB estimators for men (left panel) and women (right panel) for the period 2005-2007.

As Figure 2 shows, there are minimal differences until the age of 60 (around 1%). After this, two significant peaks emerge for the cohorts born just after the Spanish Civil War and just after World War II. Both occurrences cause a concentration of births during short time-spans that are not captured by the simplest form of the estimator (CP_UD_UB). The third peak, around the age of 90, is most likely a statistical effect due to the lack of population data for these ages, as fewer people tend to achieve this age.

To investigate issue (iii), we have analysed the impact that the different estimators have on an insurance product: in this case, a renewable year-term life insurance, with a sum insured of €100,000, for ages 50, 55, 60 and 65. To compute the premiums, we have assumed no

expenses and a discount rate of zero. The premiums are displayed in Table 1. Values for women are written between parentheses.

Table 1. Premiums payable for a renewable year-term life insurance of €100,000 for men (women). Years of study: 2005-2007.

Age	*CP_UD_UB*	*CP_NUD_UB*	*OP_NUD_NUM_UB*	*OP_NUD_NUM_NUB*
50	422.28 €	422.25 €	427.93 €	421.01 €
	(182.62 €)	(182.62 €)	(185.03 €)	(182 €)
55	637.28 €	637.21 €	645.66 €	636.66 €
	(246.14 €)	(246.14 €)	(249.07 €)	(245.84 €)
60	1,004.54 €	1,004.45 €	1,012.87 €	995.35 €
	(369.03 €)	(369 €)	(371.54 €)	(365.91 €)
65	1,430.16 €	1,429.33 €	1,447.76 €	1,410.84 €
	(562.58 €)	(562.44 €)	(570.03 €)	(554.41 €)

In general, the OP_NUD_NUM_NUB estimator premium is cheaper at ages approaching retirement for men and women. The biggest differences are 2.20% for men (2.28% for women) at age 64.

5. Conclusions and Future Research

In the actuarial and demographic fields, understanding mortality is always an important issue. The results obtained in this research have proven the impact of uniform hypothesis on both death probabilities and premium calculation. This has ramifications on best-estimate technical provisions under Solvency II and the new IFRS-17 regulatory frameworks. As demonstrated with the database analysed in this research, assuming a uniform distribution of birthdates is not appropriate. Hence, the estimator we propose in section 2.6 should be encouraged from a theoretical and practical perspective. Anyway, it would be interesting to compare for the whole range of ages all the statistical data contained in triangles BCF and DHG, for the annual estimator, and BCF and PTS, for the three-year estimator, as they have an impact on the numerator but not on the denominator of our estimators.

Acknowledgments

This research has been supported by the Spanish Ministry of Science, Innovation and Universities and the Spanish Agency of Research, co-funded with FEDER funds, project ECO2017-87245-R and Generalitat Valenciana (Consellería d'Innovació, Universitats,

Ciència i Societat Digital) project AICO/2019/053. The authors wish to thank Marie Hodkinson for revising the English text.

References

Arias, E. (2011). United States life tables. *National Vital Statistics Reports*, 64, 1-64.

Graunt, J. (1662). *Natural and Political Observations Made upon the Bills of Mortality*. London: Roycroft.

INE (2016). *Tablas de Mortalidad*, Madrid: Instituto Nacional de Estadística (Spain) [online]. Available at goo.gl/8Ywdc9.

Lexis, W. (1880). La representation graphique de la mortalité au moyen des points mortuaires. *Ánnales de Demographie Internationale*, 4, 297-324.

Lledó, J., Pavía, J.M., Morillas, F. (2019). Incorporating big microdata in life table construction: A hypothesis-free estimator. *Insurance: Mathematics and Economics*, 88, 138-150.

ONS (2012), "Guide to Calculating Interim Life Tables," Hampshire: Office for National Statistics (UK) [online]. Available at goo.gl/xFz7bV.

R Foundation for Statistical Computing (2019). Vienna. http://www.R-project.org/

Ruggles, S. (2014). Big microdata for population research. *Demography*, 69, 287-297.

Wilmoth, J. R., Andreev, K. F., Jdanov, D. A. and Glei, D. A. (2007). Methods protocol for the Human Mortality Database. *Human Mortality Database*. University of California Berkeley and Max Planck Institute for Demographic Research [online]. Available at http://www.mortality.org/.