A method for estimating the age-specific mortality pattern in limited populations of small areas

Anastasia Kostaki[1] Byron Kotzamanis[2]

1. Introduction

For many purposes in population analysis, in medical research and actuarial practice, there is a need of analytical and reliable mortality estimations. In population analysis, the age-specific mortality rates are required in order to construct complete life tables, for providing population projections differentiated by age, as well as for calculations of fertility measures, as net reproduction rates. In biomedical and epidemiological studies the age-specific mortality rates are useful for mortality comparisons between groups and for constructing multiple decrement tables. In addition, in actuarial practice the age-specific mortality rates are required for designing health and social security systems as well as for designing life insurance and pension programs. However the empirical mortality evidence is often affected by problems that not allow the calculation of reliable and accurate age-specific mortality rates. These problems can be classified into three categories. In the first category are problems related to limitations in the sense that the empirical data are either provided in an aggregated form (death counts are given in five-year or wider age groups) or they are incomplete in the sense that they are provided for some ages but not for others. A second category of problems are those related with the reliability of the empirical data. The most typical of these problems is the appearance of age heaping, i.e. at age declaration, a preference of the responder to round off the age of the descendent in ages that are multiples of five. A third category of problems are those related with the accuracy and efficiency of mortality rates when they refer to limited sized populations as for example those of small population areas in population analysis, or small target groups in biomedical analysis and actuarial proctice. It is easy to understand the nature of problems of the two first categories though not always as easy to deal with them. Usually, the use of advanced statistical tools and software are required.

[1] Department of Statistics, Athens University of Economics and Business, Greece.
kostaki@aueb.gr

[2] Department of Planning and Regional Development, University of Thessaly. Greece.
bkotz@prd.uth.gr

However problems of the third category can be sneaky in the sense that they can often be neglected by researchers or practitioners with limited statistical background. This paper provides a review of all these problems and also proposes ways out. At the outset the problems are presented one by one and ways for overcoming them are discussed and presented.

2. Data limitations

Aggregation

For several reasons the empirical death counts and therefore even the empirical death rates are provided in aggregated form (in five-year or wider age groups). Several statistical offices provide death counts differentiated in five-year age groups while for the later adult ages (above 85), they provide an aggregated death count. The reason of this limitation is the appearance of age misclassifications of the empirical death counts, the most typical of which is the appearance of age-heaping, i.e. a preference of the responders to round off ages at declaration in multiples of five. A way out of this inconsistency is to group the events in five-year age groups under the assumption that the innaccuracies in single ages into each five-year age group offset each other. In such a case the application of an expanding technique on the grouped counts and rates is appropriate in order to estimate the age-specific ones.

Incompleteness of empirical data

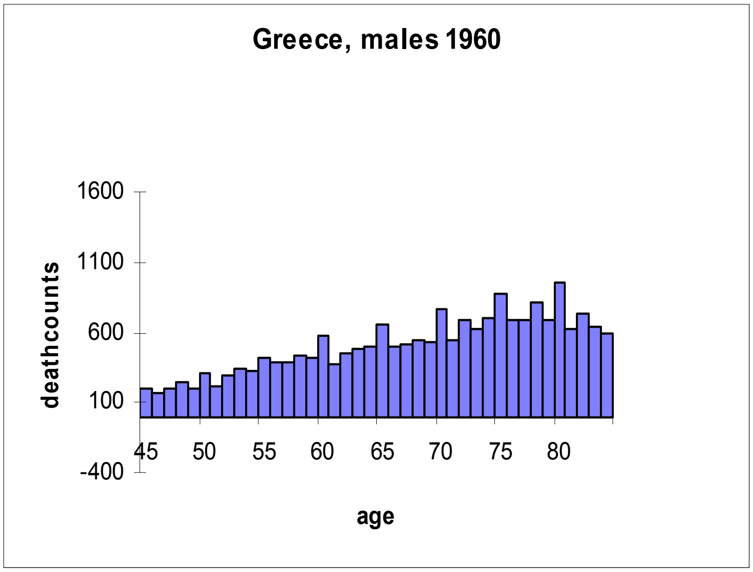
In many cases the age-specific death counts and rates are incomplete in the sense that they exist for some ages but not for others. This problem is often evident when the population groups considered are small sized. This problem is especially actual in biomedical research when the groups of patients considered are small, as well as, in actuarial analysis when the calculations are based on existed modest samples of insured persons, but also in demographic analysis when it deals with limited populations of small geographical areas.

In such a case, a possible way for estimating the age-specific probabilities of dying for each single age is to fit a parametric model (e.g. Heligman and Pollard, 1980; Kostaki,

1992) to the existed one-year empirical death rates in order to produce estimates for the missing ones. The parametric modeling has the advantage of providing smooth results, giving thus closer estimates to the true probabilities of dying underlying the empirical rates, under the assumption that the later follow a smooth pattern. An alternative way is to apply a non parametric graduation technique to the existed one-year death rates e.g. Kernels (Kostaki and Peristera, 2005) or Support Vector Mashines (Kostaki, et al., 2010).

Low reliability of empirical data

Empirical age-specific counts and rates are often affected by systematic sourses of errors, the most typical of which is the problem of age heaping mentioned above. This problem is very easy to detect in empirical death counts illustating them in a simple bar chart by age. As an example figure below illustrates the death counts by single year of age of Greek males for the year 1960. The bars for the multiples of five ages are obviously higher than for the sorrownding ones.



The prevalence of age heaping, or the tendency to over-report ages ending in 0 or 5 is usually measured using Whipple index (Shryock and Siegel 1976) calculated as

$$\frac{\sum_i D_{5i}}{\frac{k}{n}\sum_{j=x_{\min}}^{x_{\max}} D_j} * 100$$

where D_x are the death counts at age *x,* *k* is number of ages that are multiples of 5 in the age interval considered, and *n* is the total number of ages considered. This index takes the value 100 if the data are free from age heaping. In the example above the value of the index is 118.

In order to estimate the age-specific death rates in such a case, an effective way is to group the empirical death counts in five-year groups with central ages the multiples of five and then to calculate the five-year death rates for these groups. Then applying an expanding technique on these rates, estimations of the age specific ones can be provided.

Limited exposed-to-risk populations

Beside the problems mentioned above, another problem can arize if the empirical data come from small populations. This is a sneaky problem that easily can be neglected from the researcher, though its impact can lead to seriously misleading results and conclusions even if the data are complete and reliable.

Let us consider this problem, which might be highly current when we deal with spatial (small area) population analysis or limited target population samples, in demographic, medical research and actuarial practice.

Consider D_x to be the observed death count at age *x*. The death count D_x is a random variable which is binomially distributed with expected value $E(D_x) = E_x\, q_x$ and variance $Var(D_x) = E_x\, q_x\, (1 - q_x),$ where E_x is the exposed-to risk population at age *x,* while q_x is the unknown theoretical probability of dying at age *x*. Since q_x is low and E_x is large enough, D_x can also be considered as approximately Poison distributed with $E(D_x) = E_x\, q_x$ and $Var(D_x) = E_x\, q_x$. In addition according to central limit theorem D_x can also be considered as asymptotically normal distributed. Therefore the observed death rate, $\dot{q}_x = D_x / E_x$ can also be considered as asymptotically normal distributed, with $E(\dot{q}_x) = q_x$ and $Var(\dot{q}_x) = \frac{q_x \cdot (1 - q_x)}{E_x}$.

Hence, the unknown probability of dying at the age interval q_x is expected to take a value in the interval:

$$\dot{q}_x \pm 3\sqrt{\dot{q}_x * (1-\dot{q}_x) / E_x} \qquad (2.1)$$

or since $(1-\dot{q}_x)$ is near unity,

$$\dot{q}_x \pm 3\sqrt{\dot{q}_x / E_x}.$$

Let us now consider now an example. In the area of Eurytania in Greece during the five-year period 2006-2010, the exposed population at age 10 is E_x*=1888,* while the death count for the age group D_{10} *= 0,* thus the observed death rate $\dot{q}_{10}$ *= 0.* Hence, it is not possible to provide estimate for the value of the probability of dying at age 10, based on the empirical evidence.

Let us now consider a further example from the same population. The count of deaths at the age 55, $D_{55} = 6$ and the exposed-to-risk population, E_{55} *= 1464*. Thus $\dot{q}_{55}$ *= 0,0041* Putting these values in (2.1), we conclude that conclude that*: -0,0009<* $\dot{q}_{55}$ *<0,0091*. Here again $\dot{q}_{55}$ proves highly inaccurate estimator of q_{55}, and the reason is that D_{55} is too small. In general, since q_x is very small and thus (1-q_x) is near unity, the confidence interval for q_x can approximately be calculated as:

$$\frac{D_x}{E_x} \pm 3 \cdot \frac{\sqrt{D_x}}{E_x}$$

Since nobody is immortal, a minimum requirement is that the lower limit of the above confidence interval should be positive, i.e.

$$\frac{D_x}{E_x} \geq 3 \cdot \frac{\sqrt{D_x}}{E_x} \Rightarrow D_x \geq 3\sqrt{D_x} \Rightarrow D_x \geq 9.$$

Hence, independently of the population size, more than nine events are required in order to fulfill the minimum requirement that is the lower limit of the confidence interval to be positive.

A more properly defined confidence interval, according to Garthwaite, et al (2002), can be derived from the following:

As known,

$$\mathrm{P}\left[\frac{(D_x - E_x \cdot q_x)^2}{E_x \cdot q_x \cdot (1-q_x)} \le \chi^{2\;(1)}_{1-a}\right] \approx 1-a \Rightarrow$$

$$P\left[E_x\, q_x^2\,(E_x + z^2_{1-a/2}) - E_x\, q_x(2D_x + z^2_{1-a/2}) + D_x \;\le 0\right] \approx 1-a \Rightarrow$$

$$P\left[\psi(q_x) \le 0\right] \approx 1-a$$

The inequality in the probability statement is satisfied for the values of q_x which lie between the roots $q_{x,1}$, $q_{x,2}$ of the quadratic equation $\psi(q_x) = 0.$

Thus we have,

$$P\left[q_{x,1} \le q_x \le q_{x,2}\right] \approx 1-a.$$

where $q_{x,1}$, $q_{x,2}$ are given by

$$\frac{(2D_x + z^2_{1-a/2}) \pm z_{1-a/2}[z^2_{1-a/2} + 4D_x(1-(D_x / E_x))]^{1/2}}{2(D_x + z^2_{1-a/2})}$$

Considering the previous example, where *D_{55}=6* and *E_{55} =1464,* and calculating now the above confidence limits for *1-α = 0,99* we conclude that *0,0012<q_x<0,0130.* This alternative confidence interval is better than the previously considered classical one in the sense that both its limits are positive. However it is still too wide to be efficient.

There are two possible simple ways out of this serious limitation of empirical data. The first one is to group the data in five-year and the other is to consider wider periods of investigation, while if the exposed-to risk populations are too small probably both ways are required. However for several reasons in demographic and biostatistical research there is a need for reliable estimations of death counts and probabilities specific by age. For the actuary analytical and accurate mortality estimations differentiated by age are requested, having information based only on a limited sample comprising the policy holders of a Life Office.

For all the reasons mentioned above, the need for expanding the grouped data to age-specific ones should be plausible and required.

Some methods have presented in the actuarial and demographic literature for expanding death rates, e.g. the use of a six-point Lagrangean interpolation formula applied to the survival probabilities of *l(x)* of the abridged life table in order to provide estimations of these probabilities for ages that are missing from the abridged life table. Descriptions of this method are given in Elandt-Johnson and Johnson (1980), and Namboodiri (1991). Another technique for expanding an abridged life table was developed by Kostaki (1987, 1991) also provided in the MORTPAK software package (United Nations, 1988a, 1988b). In this technique the Heligman-Pollard formula (Heligman-Pollard, 1980) is utilized. Some years later Kostaki (2000) developed a nonparametric relational expanding technique much simpler than all the above with much better performance than the former while at least equal successful performance as the latter.

In addition for estimating age-specific death counts, from death counts given in five-year groups, Kostaki and Lanke (2000) developed a degrouping technique that utilizes the Gombertz classical law of mortality.

In this work a technique for estimating age-specific death counts, from death counts given in five-year groups is provided. This technique is based on the ideas of Kostaki and Lanke (2000) and utilizing the expanding technique of Kostaki (2000).

3. A technique for estimate age-specific death counts data given in age groups

Consider the empirical death count for the five-year age interval [x, x+5), ${}_5D_x$, *x=0, 5, 10,... w-5.* Then the five-year death rates ${}_5q_x$, *x*=0, 5, 10, … are calculated by

$$ {}_5q_x = {}_5d_x / \sum_{y \geq x} {}_5d_y \tag{3.1}$$

where the summation in the denominator of (3.1) is restricted to multiples of five. Obviously the consideration of the exposed-to-risk population of a given age as the sum of deaths after that age is precisely valid when the data concern a closed cohort. However, as it will be demonstrated, this procedure produces excellent results.

Our next step now is to expand the abridged ${}_5q_x$ values as calculated using (3.1). For that we consider a set of one-year probabilities, $q_x^{(s)}$ (*S* for Standard) of a standard complete life table. Under the assumption that the force of mortality, $\mu(x)$, underlying the target abridged life table is, in each age of the five-year age interval *[x, x+5),* a constant multiple of the one underlying the standard life table in the same age interval, $\mu^{(S)}(x)$, i.e

$$\mu(x) = {}_5K_x * \mu^{(S)}(x) \tag{3.2}$$

the one-year probabilities q_{x+i}, *i= 0,1,..., 4,* for each age in each five-year age interval can be calculated using

$$q_{x+i} = 1 - (1 - q_{x+i}^{(S)})^{{}_5K_x} \tag{3.3}$$

where

$${}_5K_x = \frac{\ln(1 - {}_5q_x)}{\sum_{i=0}^{4} \ln(1 - q_{x+i}^{(S)})} \tag{3.4}$$

An inherent property of the new technique is that its results fulfill the desired relation:

$$1-\prod_{i=1}^{4}\ (1-\tilde{q}_{x+i}) = {}_5q_x$$

Finally, using the estimated age-specific death probabilities, as calculated using (3.3), we provide estimations of the age-specific death counts, $\hat{d}_x$ which for the ages that are multiples of five will be calculated using

$$\hat{d}_x = \hat{q}_x \cdot \sum_{y \geq 5} {}_5d_y \qquad (3.5)$$

while for the rest of ages, $\hat{d}_x$ will be calculated using

$$\hat{d}_{x+i} = \prod_{j=0}^{i-1}(1-\hat{q}_{x+j})\ \hat{q}_{x+i} \cdot \sum_{y \geq x} {}_5d_y, \qquad i = 1,2,3,4 \qquad (3.6)$$

It is interesting to observe that the resulting $\hat{d}_x$ fulfil the property

$$\sum_{i=0}^{4} \hat{d}_{x+i} = {}_5d_x$$

The results are very close to the true values and also fulfill desirable properties. The choice of the standard table does not affect the results. The procedure is very easy to apply.

References

Garthwaite, P.H., Jolliffe, I.T., Jones, B. 2002: *Statistical Inference* 2nd edition, Oxford Science Publication.

Elandt-Johnson, R., Johnson, N. 1980: Survival Models and Data Analysis. New York, John Wiley.

Heligman, L., Pollard, J. H. 1980: The Age Pattern of Mortality. Journal of the Institute of Actuaries, 107: 49-80.

Kostaki, A. 1987: The Heligman-Pollard Formula as a Technique for Expanding an Abridged Life table. Technical report, Lusadgd-SAST-3116/1-15. Department of Statistics, University of Lund, Sweden.

Kostaki, A. 1991: The Heligman-Pollard Formula as a Tool for Expanding an Abridged Life table. Journal of Official Statistics, 7(3): 311-323.

Kostaki,A., 1992: "A Nine-Parameter Version of the Heligman-Pollard Formula". Mathematical Population Studies, 3(4), 277-288.

Kostaki A., Lanke J. 2000: 'Degrouping mortality data for the elderly" Mathematical Population Studies, 7(4), 331-341.

Kostaki A. 2000: "A relational technique for estimating the age-specific mortality pattern from grouped data". Mathematical Population Studies, 9(1),.83-95.

Kostaki, A., Panousis, E. 2001: "Methods of expanding abridged life tables: Evaluation and Comparisons" Demographic Research, 5(1), 1-15. http://www.demographic-research.org/volumes/vol5/1/

Kostaki, A., Peristera P. 2005: " Graduating mortality data using Kernel techniques: Evaluation and comparisons" Journal of Population Research, 22(2), 185-197 .

Kostaki, A., , Moguerza,M.J., Olivares, A., Psarakis, 2010: "Support Vector Machines as tools for Mortality Graduations" to appear in Canadian Studies in Population 38(3–4), 37–58.

Namboodiri N. K. 1991: Demographic analysis: A stochastic approach. Academic Press. San Diego.

Shryock, H. S., Siegel J. 1976: H, Methods and Materials of Demography. Academic Press. New York.

United Nations (1988a) MORTPAK. The United Nations Software Package for Mortality Measurement (Batch oriented Software for the Mainframe Computer), New York: United Nations: 114-118.

United Nations (1988b) MORTPAK-LITE The United Nations Software Package for Mortality Measurement (Interactive Software for the IBM-Pc and Compatibles), New York: United Nations: 111-114.