Estimating probability distribution for age at death
Bayesian method
User's guide to R programs

Henri CAUSSINUS and Daniel COURGEAU

Disclaimer. *The programs presented below are tools that, in their present state of development, are not designed to use all the resources of R software. They should be seen as provisional and subject to further improvements.*

The **Estimation1** program uses a D irichlet prior distribution for the probabilities of age classes.

The **Estimation2** program uses two different priors: the Dirichlet and a uniform distribution on a set of candidate vectors. This user's guide addresses first the **Estimation1** program. It then provides the modifications that concern the use of the **Estimation2** program.

Estimation1

Objective and results provided

Starting with the distribution of bone stages at a target site and reference data, the program estimates the probabilities for various age classes using the Bayesian method recommended by Caussinus and Courgeau in Section B of the *Manuel de Paléodémographie* (see also Caussinus and Courgeau, *Population,* 2010). The method is non-parametric in the sense that it does not use an explicit mortality model. The calculations may, however, be supplemented by adjusting a mortality distribution (Gompertz or Gompertz-Makeham).

The program produces

1) Parameters for the posterior distribution: means, standard deviations, correlation coefficient matrix, quantiles for each marginal distribution;

2) Posterior cumulative distribution function graphs for each age class, and a posterior densities;

3) Adjustment of a Gompertz distribution and a Gompertz-Makeham distribution: estimation of parameters and graphical display.

Statistical data

Ages and bone stages are divided into classes. The frequencies of bone stages at a site are given as a frequency vector and reference data are arranged in a frequency matrix (stages in rows, ages in columns).

The reference data are to be read from a text file (separated by spaces). In the supplementary material to the programs there are a number of files taken from the Lisbon reference population for various numbers of stages and ages, according to one or both sexes (e.g., Lisb5x8H for 5 stages, 8 age classes, for men only). Any other table can be reconstituted from the detailed data provided in this CD-ROM (files.txt).

Enter file name in command

ref<-read.table ("file name.txt").

Enter site data from the keyboard in the form

sit<-c(succession of frequencies with decimal points, separated by commas)

Enter the limits of the age classes in the form

a0<-lower limit (in years)

ac<-c(following limits in years, separated by commas)

Important: no higher limit for the final class, which is taken to be open-ended.

(Note: the class limits are not used in the calculations, except for the adjustment of a mortality distribution, but they are required for the presentation of the results.)

Tuning parameters

Enter the parameters of the Dirichlet prior distribution in the form

bet<-c(succession of values with decimal points, separated by commas)

For the other data, the program's default values may be modified, but in general this will serve little purpose.

They are

ns = number of repetitions for evaluating integrals by the Monte Carlo method (default value 100 000)

coeff = reduction coefficient applied to reference data (default value 1).

na = number of x-axis points for discretising posterior cumulative distribution functions (default value 1000).

nad = number of x-axis points for discretising posterior densities (default value 100; nad<na is recommended because the densities are used for graphical displays and it is important to produce a sufficiently smoothed version).

q = probability vector corresponding to the quantiles one wishes to express (default value q=c(0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95)).

init = initial values for the iterative process used in adjusting a Gompertz distribution (default value init= c(0.0086, 0.067)).

Program structure

The program is structured as follows:

- reading/entering data and parameters; console check
- estimation of the posterior distribution – basic calculations
- graphical displays of prior and posterior densities
- adjustment of a Gompertz distribution and a Gompertz-Makeham distribution.

Results

The numerical results directly produced are

- means and standard deviations of the posterior distribution for each age class
- correlation coefficient matrix for the posterior distribution
- quantiles of the posterior distribution for each age class

The results are displayed on the console and recorded in a text file entitled “resultats” (“results”).

The graphical displays compare prior densities (red) and posterior densities (black) for each age class; vertical lines from the x-axis indicate prior (red) and posterior (black) mean values.

By default the program displays the graphs for 9 age classes at most. The displayed graph can be saved. To display fewer classes at once (or more, but this is not advised) the layout needs to be modified at

layout (matrix(1:number of classes displayed, number of rows, number of columns, byrow=TRUE))

To display classes m to n, replace in the following row (h in 1:7) by (h in m:n); to display a selection of classes, say 1,4,6,8,9, replace (h in 1:7) by (h in c(1,4,6,8,9)).

The graphs can be saved: in the File menu, select “save as” and choose save format. Note that it is often preferable to save a figure after enlarging the window in which it is displayed.

Important:

- do not include a class number higher than the total number of classes in the example shown;
- the figures must be displayed (and saved if desired) one by one, which means running the program stage by stage when producing graphs.

The program will fit a Gompertz distribution and a Gompertz-Makeham distribution to the estimated probability of death by age classes (see Appendix B). It produces:

- estimated values of the distribution parameters, and the residual square deviation
- a graph comparing probabilities of survival estimated by the Bayesian method (circles for class limits), probabilities of survival according to the adjusted Gompertz distribution (red line) and probabilities of survival according to the adjusted Gompertz-Makeham distribution (green line).

Estimation2

The **Estimation2** program produces the same results as **Estimation1** for a Dirichlet prior distribution, plus similar results with a uniform prior distribution on a set of "candidate" vectors. The set is based on the Bocquet-Appel and Bacro method; we recommend using the sets given on Bocquet-Appel's website whenever the conditions of application permit. Furthermore, the program provides some evidence for comparing the two ways of using the recommended Bayesian method. Below we indicate only the modifications to the **Estimation1** program. Methodological guidance for choosing the prior distribution can be found in Appendix C.

Data

The candidate vectors come from a text file (separated by spaces) and are loaded with command

vp<read.table("file name.txt")

(the program is designed to directly use the files available on Bocquet-Appel's website: that is the purpose of instruction vp<-t(vp[,1:7]), which will need to be modified for other file types).

Results

Results are produced on the same principle as for **Estimation1** (console and saving in the "results" file, graphs), duplicated because the results are provided for each of the prior distributions. Furthermore, comparative graphs can be obtained: comparisons by age classes of the two prior densities and the two posterior densities.

Note. The uniform prior distribution, and consequently the corresponding posterior distribution, are discrete distributions: the densities do not have the same meaning as with the previous distributions (continuous); so it is not exactly the densities that are given by the graphs but rather curves that are intended to be smoothed representations of them.

Appendix A. R software: principles of use

R is an open-source software package that can be downloaded from a number of websites, such as

http://cran.cict.fr

R can be used with Windows, MacOS X and Linux. This is how to use one of the programs described in this paper with Windows (for MacOS X and Linux see website above).

- Download R.2.10.1 for Windows from the site to the ***Program Files*** folder, which creates a folder entitled R and places an R icon on the desktop and the Start menu.
- Create a folder to be the "current directory" for the programs and files used (particularly the reference data files for all the programs and the candidate vector files for Estimation2).
- Click on the R icon. The "console" appears.
- Go to File menu and select "change current directory", so as to go to the directory prepared for that purpose.
- Go back to File menu and select "open a script": this opens the directory where the programs are; select desired program. A new window opens, containing the program.
- To run part of the program, select it with the mouse and click on the third icon from the top left (window shape with arrow). This runs the part of the program selected; it scrolls in the console, which also displays the results obtained if results are included in this part of the program.

Notes

In the program, content between straight single quotation marks 'thus' is not executed; these are comments.

To begin with, it is advised to run the program part by part, using the comments; for example, start by going as far as 'vérification des données' ('data check'): these appear on the console, etc. This partial run is essential for graphs (see above).

The results can be read on the console and saved in a "results" file, called "résultats", if desired (by running the relevant part of the program). The file will be located in the current directory; it can be read (and retrieved) by going to that directory (always via the File menu), but in that case, retrieve "all the files" because by default only the programs appear. Important: the "résultats" file is designed to contain the successive results for a given application and that one only; a new application deletes the results previously saved; the simplest thing is to store them in a file with another name or located outside the current directory.

Appendix B. Estimating the Gompertz distribution or Gompertz-Makeham distribution closest to the observed distribution

1. Gompertz distribution

In 1825, Benjamin Gompertz, on the basis of data from various countries, proposed the distribution that now bears his name, often applicable to adult mortality at least as far as the age of 80. It represents the instant mortality rate by the formula

$$h(x) = -\frac{1}{S(x)}\frac{\partial S(x)}{\partial x} = \lambda\rho e^{\rho x}$$

where $S(x)$ is the probability of surviving from birth to age *x*, and λ and ρ are two parameters to be estimated. Parameter λ is the general level of mortality and parameter ρ the "deterioration or an increased inability to withstand destruction" with age. The differential equation may be integrated thus

$$\ln[S(x)] = -\int_{a=0}^{x} \lambda\rho e^{\rho a} da = \lambda\left(1 - e^{\rho x}\right)$$

The cumulative death rate between ages *x* and *x+n* may be expressed as

$${}_nH_x = \int_{\xi=x}^{x+n} \lambda\ \rho e^{\rho\xi} d\xi = \lambda\left(e^{\rho(x+n)} - e^{\rho x}\right) = \lambda e^{\rho x}\left(e^{\rho n} - 1\right)$$

It can be seen that this cumulative rate follows an exponential distribution of age *x* like the instant rate but with a different λ coefficient, while the ρ coefficient remains the same. But this cumulative rate differs from the standard multi-year rate.

Multi-year rates may be obtained from instant rates by the formula

$${}_nq_x = 1 - \exp\left(-\int_{\zeta=x}^{x+n} h(\zeta) d\zeta\right)$$

where *h* is the instantaneous rate.

With a Gompertz distribution, the result is:

$${}_nq_x = 1 - \exp\left(-\int_{\xi=x}^{x+n} \lambda\rho e^{\rho\xi} d\xi\right) = 1 - \exp\left(\lambda\left\{e^{\rho x} - e^{\rho(x+n)}\right\}\right).$$

Consequently we may express the probability at age *x* of surviving at least until age *x+n* thus:

$${}_np_x = 1 - {}_nq_x = \exp\left(\lambda\left\{e^{\rho x} - e^{\rho(x+n)}\right\}\right).$$

It can be seen that it is the natural logarithm of this probability of survival that also follows the Gompertz distribution, the same as for the cumulative rate.

In 1860, William Makeham, considering that allowance must be made for the fact that some causes of death, such as accidents, are independent of age, proposed supplementing the Gompertz distribution with a constant coefficient λ_0:

$$h(x) = \lambda_0 + \lambda\rho e^{\rho x}$$

.

We see that in this case, by calculating the probability at age x of surviving at least until age $x+n$

$$ {}_n p_x = \exp\left(\lambda_0 n + \lambda e^{\rho x}\left\{ e^{\rho n} - 1\right\}\right), $$

the logarithm of this probability of survival also follows a Gompertz-Makeham distribution with parameters different from the previous one.

2. Estimation

The Bayesian method provides an estimation of the probabilities for the various age classes by an a posteriori distribution. Let μ be the vector (column) of the estimated means of probabilities of survival and V the matrix of variances and covariances of this distribution.

Let $G(\lambda,\rho)$ be the vector (column) of probabilities of classes on the hypothesis of a Gompertz distribution of parameters λ and ρ (conditional on survival beyond the lowest age considered).

A simple manner of fitting a Gompertz distribution, i.e. of determining the parameters λ and ρ that best fit the observations, is to use least squares by minimising the values of λ and ρ in the expression

$$ (\mu - G(\lambda, \rho))' \, (\mu - G(\lambda, \rho)) \qquad (1) $$

It may seem more natural to use appropriately weighted least squares and minimise the values of λ and ρ in the expression

$$ (\mu - G(\lambda, \rho))' \, V^{-1} \, (\mu - G(\lambda, \rho)) \qquad (2) $$

In practice, one class in G has to be omitted so that V may not be necessarily singular; we also use cumulative probabilities to simplify the expression of G by making the appropriate transformations of μ and V (this does not alter the estimation of λ and ρ because of the linearity of the transformation). So we obtain for the c-1 elements in vector:

$$ G(\lambda,\rho) = \exp\left(\lambda\left\{ e^{\rho x_0} - e^{\rho x} \right\}\right), $$

where x_0 is the lower limit of the first age class and x successively takes the value of the c-1 lower limits of the following age classes.

The minimum of (1) or (2) is easily obtained by standard procedures. The programs use the constrOptim procedure in R in order to impose a positivity constraint on λ and ρ.

The minimum of (1) or (2) is easily obtained by standard procedures. The programs use the constrOptim procedure of R in order to impose a positivity constraint on λ and ρ. Only one difficulty occurred in practice: although the reduction to c-1 elements mentioned above does prevent V being singular, the matrix may come close to singularity for a uniform prior distribution on a set of vectors, leading to unsatisfactory results when minimising (2). For that reason, the **Estimation2** program uses by default a minimisation of (1) for a uniform prior distribution. However, it is easy to get it to minimise (2): put in straight single quotation marks the instruction 'vinvU<-diag (nc-1)' in line 11 of the 'lois de Gompertz et Gompertz-Makeham ajustées' part of the program.

To adjust a Gompertz-Makeham distribution, the procedure is similar, replacing $G(\lambda,\rho)$ by the class probability vector on the hypothesis of the new distribution $G(\lambda_0, \lambda, \rho) = \exp\left(\lambda_0\{x - x_0\} + \lambda\left\{ e^{\rho x_0} - e^{\rho x} \right\}\right)$.

Appendix C. Selecting the prior distribution

Some of the criteria for selecting the prior distribution are discussed in Chapter B of the *Manuel de Paléodémographie* and in the *Population* article (Caussinus and Courgeau, 2010). In particular there are arguments for choosing parameters for the Dirichlet prior. We shall not return to these matters. We supplement them here with some considerations concerning the choice between a Dirichlet distribution and a uniform distribution on a set of vectors.

We have shown in the *Manuel de Paléodémographie,* chapter B, that the Bayesian method with an appropriate Dirichlet distribution is overwhelmingly superior to earlier methods of estimation; in the *Population* article we observe that the only non-parametric method likely to compete with ours is that of Bocquet-Appel and Bacro; consequently we compare our method with theirs by assuming that the vector of probability to be estimated is one of their "candidate" vectors. We show that the most effective way of proceeding is to use the Bayesian method with a uniform prior distribution. This supports our method rather than Bocquet-Appel and Bacro's, since we had adopted the case which seemed on the face of it the most favourable to theirs; this also gives an advantage to the uniform prior distribution over the Dirichlet distribution (with parameters deduced from the pre-industrial standard) when the vector to be estimated is one of the candidate vectors. If we closely examine the implications of such a distribution for the prior distribution of the probability of each age class, we see that the classes are only slightly dispersed compared with the dispersion in the Dirichlet prior distributions[1]. Consequently, if we may state that the "prior" knowledge of the site considered is sufficiently accurate and compatible with the uniform prior distribution, this distribution is reasonable; otherwise, we believe it is better to stick to a Dirichlet distribution with a much lower prior impact on the final estimate. In other words, although posterior distributions are generally less dispersed with a uniform prior than a Dirichlet one, the fear is that this is a misleading precision based on an unjustified prior distribution. The graphs included in the program comparing the two types of prior and posterior distributions may enlighten the user.

[1] For example, with the candidate vector file "ProbAtri20-90.txt" the prior distribution for the 50-59 age class is concentrated between 0.112 and 0.233, which makes it impossible to make estimation outside this interval (either by our method with uniform prior or by Bocquet-Appel and Bacro's).