Chapter 13

A new method for estimating age-at-death structure

Henri Caussinus and Daniel Courgeau

13.1 Introduction

The previous chapter presents the main methods hitherto recommended for estimating a population's age structure. This chapter proposes a new method based on a precise statistical model taking into consideration the essential specificity of the data upon which the estimation is based.

First the notation, which is basically that of the previous chapter. We denote p_{ij} the probability that an individual taken at random from the study population belongs to age class j ($j = 1, \ldots, c$) and stage i ($i = 1, \ldots, l$) of a given indicator; the sum of the p_{ij} with respect to i is denoted p_j (the probability that an individual is aged j), the sum of the p_{ij} with respect to j is denoted π_i (the probability that an individual is at stage i); the conditional probability of stage i being at age j is denoted $p_{i|j}$. These various probabilities are positive and satisfy the equations $\sum_i \pi_i = \sum_j p_j = 1$ and $\sum_i p_{i|j} = 1$ for all values of j. They are also connected by the following equation:

$$\sum_j p_i \, p_{i|j} = \pi_i \text{ for all values of } i = 1, \ldots, l \qquad [13.1]$$

In practice, the estimation must be made with data n_{ij}, the number of observations of stage i and age j in a reference base ($i = 1, \ldots, l$ and $j = 1, \ldots, c$), and m_i, the number of observations of stage i at the site in question ($i = 1, \ldots, l$; $\sum_i m_i = m$): these data are shown in Tables 12.5 and 12.6 in the previous chapter. The invariance hypothesis assumes that the probability $p_{i|j}$ of stage i occurring at age j is the same for any population; it is possible therefore to calculate these conditional probabilities from the reference data even if they come from another population. Consequently, the model is parametered by $p_{i|j}$ and p_j, with the π_i being deduced if necessary from equation [13.1]; the parameters of interest are clearly the p_j, whereas the $p_{i|j}$ are only of intermediate value.

The various proposals mentioned in the previous chapter do not take fully into consideration the variability of some of the observations: for example, the IALK method replaces each $p_{i|j}$ by $n_{ij}/n_{.j}$ as if this quantity were fixed and not random[1], and Bocquet-Appel and Bacro's method (2008) does not fully consider the random nature of the m_i, since the results of the estimation depend solely on the m_i/m ratios (it is clear that the size of sample m affects the

[1] The IALK method can be modified to take account of this randomness. One proposal is given in Appendix B; while this does improve the method in some ways, it may weaken it in others, which confirms our view that more radical changes in viewpoint are necessary.

precision of the estimates, i.e., the confidence intervals). Furthermore, except for Bocquet-Appel and Bacro's method, the procedures proposed do not address the specific nature of the problem at hand. One does not estimate just any set of probabilities but rather a distribution of ages at death for a group of individuals about whom one may even have more precise specific information. If we also consider that the data available are usually scarce, this specific information is all the more valuable. This may also be concluded empirically from the high instability of most of the methods proposed up to now.

All these points have convinced us of the utility of introducing the Bayesian method we present in Section 13.2. A few simulations in Section 13.3 show that this method appears to effectively replace most of the former methods; they are also used in the discussion of certain questions of calibration. Section 13.4 addresses particular examples and provides comparisons with other approaches from a new angle.

One final point concerns vocabulary. Some methods in the literature may appear to be Bayesian in so far as they make use of the so-called Bayes formula or introduce a priori considerations into the resolution of estimation problems, but the paradigm on which they are based remains frequentist.[2] The method we present here, on the other hand, is Bayesian in the sense most often used in statistics: it is considered that the parameters themselves are random, with a probability distribution, called *prior*, chosen by the user to reflect his(her) knowledge (and ignorance) before the observation; this distribution is then corrected in response to the observations to achieve a *posterior* distribution, which is the observation-based probability distribution of the parameters, and, more specifically in our case, the posterior distribution of the parameters of interest p_j ($j = 1, \ldots, c$).

13.2. A Bayesian estimation method

13.2.1. Model and principle

It is natural to suppose that the frequencies m_i ($i = 1, \ldots, l$) observed on the site for various stages are the observed values of a multinomial distribution whose parameters π_i are linked to the p_j and $p_{i|j}$ according to equation [1]. We shall use these parameters to pursue the modelling.

We denote by G the prior density of parameters $p_{i|j}$, $i = 1, \ldots, l$ and $j = 1, \ldots, c$ (we shall see how G can be expressed in Section 2.2) and assume that the parameters p_j ($j = 1, \ldots, c$) have a prior density g (also discussed in Section 2.2) and are independent of the $p_{i|j}$.

If we denote by M the vector of m_i, P the vector of $p_{i|j}$ and p the vector of p_j, the joint density of (M, P, p) will be f given by

$$f(M, P, p) = g(p)G(P)\frac{m!}{\prod_i m_i!}\prod_i\left(\sum_j p_j p_{i|j}\right)^{m_i}$$

[2] This is true of Bocquet-Appel and Bacro's method (2008), which takes account of the nature of the probabilities to be estimated by reducing the parametric space of the standard framework.

where the index i always goes from 1 to l and the index j from 1 to c.

The marginal density of the pair (M, p) is

$$\int f(M,P,p)dP$$

and the marginal density of M is

$$\iint f(M,P,p)dp\,dP$$

whereby the integrals are taken over the variation domains of P or p and P, which are a simplex (for p) or a product of simplexes (for P).

The conditional density of p, given M, is therefore

$$\frac{\int f(M,P,p)dP}{\iint f(M,P,p)dp\,dP}$$

This is the *posterior* density of p_j ($j = 1, \ldots, c$) on which the Bayesian estimation will be based.

For example, one may have the posterior mean of p_j

$$\frac{\iint p_j\ f(M,P,p)dp\,dP}{\iint f(M,P,p)dp\,dP}$$

More generally, the conditional expectation given M of a function φ of p will be given by

$$\frac{\iint \varphi(p) f(M,P,p)dp\,dP}{\iint f(M,P,p)dp\,dP} \qquad [13.2]$$

We thus obtain, for example, the kth-order moment of p_j with $\varphi(p) = p_j^k$. Taking for $\varphi(p)$ the function that equals 0 for $p_j > x$ and 1 for $p_j < x$ (indicator variable of the event $p_j < x$), we express the posterior distribution function for p_j at point x.

The various integrals in expression [2] may be evaluated by a Monte Carlo method as follows.

We denote $X = (X_1, \ldots, X_c)$ a random vector with density distribution g and Y a family of c vectors $Y_j = (Y_{1j}, \cdots, Y_{lj})$ ($j = 1,\ldots,c$), whose joint distribution is independent of X and admits density G. We verify that expression [2] equals

$$\frac{E\left(\varphi(X)\prod_i\left[\sum_j X_j Y_{ij}\right]^{m_i}\right)}{E\left(\prod_i\left[\sum_j X_j Y_{ij}\right]^{m_i}\right)}$$

Let us generate S independent sets of such random vectors (X,Y), with s ($s = 1,\ldots,S$) representing the iterations. By virtue of the law of large numbers, if S is large enough, the expression above is approximated by

$$\frac{\sum_{s=1}^{S}\varphi(X_s)\prod_i\left(\sum_j X_{js}Y_{ijs}\right)^{m_i}}{\sum_{s=1}^{S}\prod_i\left(\sum_j X_{js}Y_{ijs}\right)^{m_i}}$$

This supplies the posterior expectation of each p_h (h = 1, … ,c) – which can be taken as a point estimate – or the posterior variance useful for characterising the accuracy of the estimate. The same principle can be applied to evaluate cross-moments, such as the covariance matrix of the posterior distribution of the p_h parameters. The posterior distribution function of a p_h parameter can be used, for example, to calculate intervals containing that p_h with a given probability, known as credibility intervals, which are the Bayesian equivalent of the confidence intervals of the standard system.

13.2.2. Use in practice

13.2.2.1. Choice of prior distributions

a. Density *G*

The only source of information on the conditional probabilities $p_{i|j}$ is the reference data. If they are raw data merely obtained by recording the stage frequencies on a sample of skeletons of known ages, we can logically conclude that, for each age class j (j = 1,…,c), the frequencies n_{ij} are the observed values of a multinomial distribution with a total n_j and probabilities $p_{i|j}$(i = 1,…,l). Adopting a prior distribution for the $p_{i|j}$ probabilities, we deduce the conditional distribution given the reference data. We take it, in turn, as the prior distribution of the $p_{i|j}$ probabilities in the final model. Given the absence of supplementary information on these $p_{i|j}$ probabilities beyond what is contained in the reference data, it makes sense to adopt a uniform distribution as the prior distribution of the $p_{i|j}$ probabilities for each j. For a given j, we find a posterior distribution of the $p_{i|j}$ probabilities that represents a Dirichlet distribution (see Box 9) of parameters $\alpha_{ij} = n_{ij} + 1$ (i = 1,…,l). Density G is then the product of c Dirichlet densities, namely

$$G(p) = \frac{\prod_j \Gamma(\alpha_{.j})}{\prod_i\prod_j \Gamma(\alpha_{ij})}\prod_i\prod_j p_{i|j}^{\alpha_{ij}-1}$$

In practice, the raw data may be "processed" in various ways (for example, in order to achieve the right weighting between male and female samples), so that their distribution is no longer strictly multinomial. However, the prior G as defined above appears still to hold, since the multinomial nature of the reference data is more an indication than a necessity for arriving at that distribution.

The choice of G may be refined in various ways. For example, in order to avoid excessive confidence in the reference data, the α_{ij} may be multiplied by a "reducing" coefficient r ($0 < r < 1$) with a choice of $\alpha_{ij} = r(n_{ij} + 1)$, which does not affect the prior mean values of $p_{i|j}$ but increases the prior variances, thereby expressing the degree of doubt. These variances are roughly multiplied by $\frac{1}{r}$; note that it is very broadly equivalent to assume that the n_{ij} are

multiplied by *r*, another way of reducing the information contained in the reference data since it amounts to assuming that the relative frequencies observed in the data reference are retained but taken from a smaller sample.

Box 13.1. The Dirichlet distribution

Let D be the subset of $\Re^k$ defined by:

$$x = (x_1, \ldots, x_k) \in D \quad \Leftrightarrow \quad x_i > 0 \quad for\ all \quad i = 1, \ldots, k \quad and \quad \sum_{i=1}^{k} x_i = 1.$$

and $a = (a_1, \ldots, a_k)$ a vector of strictly positive real numbers.

The random vector $X = (X_1, \ldots, X_k)$ follows a Dirichlet distribution with parameter a if its probability density d is such that:

$$d(x) = \begin{cases} \dfrac{\Gamma(a_.)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} x_i^{a_i - 1} & for\ x \in D \\ 0 & for\ x \notin D \end{cases}$$

where $a_. = \sum_{i=1}^{k} a_i$ and Γ is the Euler's Gamma function defined by $\Gamma(p) = \int_0^{\infty} e^{-x} x^{p-1} dx$.

Note that d is constant over D (uniform distribution) when $a_i = 1$ for all i. The marginal distribution of X_i is Beta with parameters $(a_i, a_. - a_i)$. The moments of X_i are:

$$E(X_i) = \frac{a_i}{a_.}$$

$$E(X_i^2) = \frac{a_i (a_i + 1)}{a_. (a_. + 1)} \quad , \quad Var(X_i) = \frac{a_i (a_. - a_i)}{a_.^2 (a_. + 1)}$$

More generally, the moment of order h $(h \geq 1)$ is:

$$E(X_i^h) = \prod_{j=0}^{h-1} \frac{a_i + j}{a_. + j}$$

We have also, for $i \neq j$: $E(X_i X_j) = \dfrac{a_i a_j}{a_. (a_. + 1)}$ *and* $Cov(X_i X_j) = -\dfrac{a_i a_j}{a_.^2 (a_. + 1)}$.

Remarks

1. The X_i means, $e_i = \dfrac{a_i}{a_.}$, are proportional to the a_i; they remain unchanged if the a_i are all multiplied by the same positive number s. We can write: $Var(X_i) = \dfrac{e_i (1 - e_i)}{a_. + 1}$;

hence, for equal means, the variances are larger when the a_i are smaller: if a_i is multiplied by s, then $Var(X_i)$ is multiplied by $\frac{1+a_.}{1+s\,a_.}$, a decreasing function of s (if $a_.$ is large enough, this is almost equivalent to multiplying $Var(X_i)$ by $\frac{1}{s}$).

2. In the Bayesian statistical framework, one says that the Dirichlet distribution is conjugate to the multinomial one; when the prior distribution of the parameters of a multinomial law is Dirichlet, the posterior distribution is another Dirichlet distribution whose parameters are obtained by adding the vector of observed numbers to the parameter vector of the prior.

b. Density *g*

The choice of the prior distribution for the p_j parameters is a trickier matter. We give our preferred method first, which will be systematically used in this chapter. But we shall also briefly mention other possibilities, some of which merit further examination.

As there is no clearly designated "class" of distributions from which to select the prior distribution, the most sensible course is to opt for a Dirichlet distribution, which is well suited to probability vectors. This leaves the problem of choosing the distribution parameters, say ($\beta_1,\ldots,\beta_c$). In the absence of specific information, we can, as above, choose a uniform distribution and take $\beta_j = 1$ for all j. This is a "neutral" choice and may sometimes be justified. It also yields reasonable results with simple examples. However, in paleodemography, other choices would appear to be preferable as certain information is naturally available. We can, for example, take a "standard" mortality distribution and calculate the probabilities for each of its age classes. The class probabilities become the means of the prior distribution. This gives the parameters β_j up to a proportionality coefficient (see Box 9), i.e. the $\beta_j / \beta_.$ values, where $\beta_.$ is the sum of the β_j parameters over $j = 1,\ldots,c$. The remaining step is to choose $\beta_.$, i.e. in practice, the prior variances. Note that the variances need to be relatively large in order to express the fact that the prior means are not very reliable and that the prior distribution should not play a dominant role – in other words, that the family of possibilities envisaged covers a broad field. Hence $\beta_.$ should be fairly small, say, below unity or barely above. We shall see that this is indeed the case in the simulations examined below.

Note that the prior means may be seen as "test" values: if the data are scarce and the estimates consequently imprecise, it is helpful to use the posterior distribution qualitatively by observing in which direction these means move, i.e. how the data "correct" the prior values.

This principle for the choice of the prior distribution may be extended in a number of ways. For example, instead of choosing a standard mortality distribution as the basis for constructing the prior distribution, one may choose a mix of two "standard" distributions, leading to a mix of two Dirichlet distributions. These might be the mix (in carefully chosen proportions) of a routine mortality distribution (attrition) and a catastrophic distribution.

Clearly, quite different approaches are also possible, such as, along the lines of Bocquet-Appel and Bacro's proposals (2008), defining the prior distribution as a uniform distribution on a finite set of distributions corresponding to standard mortality distributions. We have examined this in Caussinus and Courgeau (2010), together with the comparison of our method with that of Bocquet-Appel and Bacro. As standard mortality distributions, rather than the "artificial" distributions proposed by these authors, one may consider the

distributions used to construct the pre-industrial standard. This deserves further research; however, this type of prior distribution is likely to place too much weight on routine mortality and be less effective in identifying specific situations of interest. Even if a Dirichlet prior as described is not necessarily optimal, it does provide a flexible general approach that is easy to implement: it is therefore the only one we consider below.

13.2.2.2. Posterior distribution and credibility intervals

Earlier, we saw how to calculate the posterior distribution function for each p_j point by point. The posterior density for each p_j can be numerically derived and then appropriately smoothed. A graphical display of the densities may help in interpreting the numerical results. In some cases, the posterior density of each p_j can be approximated by a Beta density with the same mean and variance, which, for example, simplifies the evaluation of densities. Approximation quality can be controlled to a certain extent via higher-order moments: one can check the proximity of the beta distribution's third- and fourth-order moments with respect to the corresponding moments of the "true" posterior distribution, easily calculable by simulation, as seen above. Note, however, that this type of approximation is not always valid and must be used with care, and avoided in those cases where exact calculations can readily be performed.

After calculating the posterior distribution function for each p_j, we can determine *α-credible intervals* (Robert, 2006, p. 278) in which a p_j parameter has a probability $1 - \alpha$ conditional upon the observations. It is preferable to use the exact posterior distribution function, but in some cases the approximation by beta distribution mentioned above[3] is acceptable.

Finally, note that it is extremely inadvisable to use an interval of the "mean plus or minus one (or two) standard deviations" type because the posterior distribution is, in most cases, highly asymmetric.

13.2.2.3. Size of the reference data table

System [13.1] described in Section 13.1 is undetermined if the number of rows (stages) l is smaller than the number of columns c (ages). In other words, the parameters of interest are not identifiable, given that several values lead to the same distribution of observable samples. The Bayesian method avoids the difficulty by starting with a prior distribution, and the aim is simply to make it change by means of the data. The posterior distribution steers us towards a distribution of the unknown parameters, which is wholly compatible with the fact that they are not completely determined. This method can therefore be used with $l < c$. Clearly, the posterior distribution can be somewhat dispersed, which merely reflects the indeterminacy inherent in the situation.

13.3. Brief simulation study

The following examples, taken from a wider study, are intended to illustrate the properties of the recommended Bayesian method and to specify certain points in the choice of parameters for the prior distributions. The first two examples are elementary and do not refer specifically to the nature of the underlying application, although they are described in the "language of paleodemography" (ages, stages) for the sake of consistency and to simplify explanation. The third and fourth examples are more directly connected to applications in paleodemography. In order to compare the Bayesian method with frequentist methods, we only consider the point estimates it provides with the posterior mean.

---

[3] One example is the set of data processed in Section 4. But there are cases where this approximation is highly unsatisfactory: an example of a bimodal posterior distribution is even given in Séguy, Caussinus, Courgeau and Buchet (2012).

In all events, we start from a population distributed by two discrete features with l (in lines) and c (in columns) classes respectively. We assume as known the probabilities p_{ij} that an individual will be located in line i and column j. We then simulate a large number R of situations, taking each time a multinomial sample with l categories and probabilities π_i in order to simulate the site data and also c multinomial samples with, for the jth, the probability $p_{i|j}$ for the ith class, in order to simulate the reference data. Each repetition leads to estimates of the probabilities p_j (the desired structure by age) by various methods.

In Examples 1, 2 and 3, we evaluate the least-squares regression method, the IALK method (which we prefer to call "Maximum Likelihood 1"), the Maximum Likelihood 2 method described in Box 10, and the proposed Bayesian method, by comparing the estimates found with the true values of the parameters, known in this case. These results are given both graphically in the form of frequency histograms for the estimated probability of being in one of the age groups, and in the form of standard summaries: mean, standard deviation and mean squared error. We know that the mean squared error of an estimator X of the real parameter θ equals the expected value of the square of the X-θ difference, or $E[(X-\theta)^2] = Var(X) + E[(X-E(X))^2]$; it accounts therefore both for the variance of the estimator, the first term in the sum above, and for its bias, the second term in the sum.

In Examples 3 and 4, we add the comparison with the Bocquet-Appel and Bacro method (2008), since its restriction of parametric space only becomes fully meaningful with paleodemographic data; in both examples we have chosen a breakdown into age classes compatible with the authors' "prior" datasets. In these larger examples, the quality of results is examined with an overall criterion of distance between the vector of true probabilities and the vector of estimated probabilities. In fact, two criteria are used: the sum of mean squared errors obtained for the various age classes ("total MSE") and an analogous sum weighted by the true probabilities [as in a chi-square test], ("relative MSE").

Example 1

We first take the two-row two-column example from the previous chapter, drawing multinomial samples of 20 by the probabilities considered in that chapter: as conditional probability (reference) for Line 1 we have 0.667 for Column 1 and 0.25 for Column 2; the marginal probability for Line 1 is 0.6. There is only a single parameter to be estimated in this case, for example, the probability for Column 1, which we know to be 0.84. We run 1,000 iterations, obtaining different samples. Since l = c, equivalent results are obtained for the "corrected" regression (i.e., least squares subject to positivity constraint) and Maximum Likelihood 1 (IALK).

"Ordinary" regressions with no positivity constraint run on each of these samples gave 347 "estimates" greater than unity, which is understandable since the true probability is fairly close to unity, and also 11 negative values, which is more surprising. In fact, there is a wide dispersion of results (standard deviation 0.63) around a mean close to the true value, but which hardly make sense. Correcting the higher estimates to unity and the negative ones to zero, the mean obtained is 0.78, standard deviation 0.25, and mean squared error 0.06. The histogram of estimated values is given in Figure 69 (left).

Box 10. Another maximum likelihood approach: ML2

Let us consider the following statistical model which takes into account the random character of all the data. The set of parameters is the set of probabilities p_{ij} (i = 1, …, l and j = 1, …,

c); their sums over *i* are denoted p_j while the sums over *j* are denoted π_i. For each j $(j = 1,\ldots,c)$, the n_{ij} $(i = 1,\ldots,l)$ are the observed frequencies of a multinomial distribution with l categories, the total number of trials being n_j and the cell probabilities $p_{i/j} = p_{ij}/p_j$ $(i = 1,\ldots,l)$; the frequencies m_i *(i=1,...l)* follow a multinomial distribution with l categories, total number of trials m and cell probabilities π_i $(i = 1,\ldots,l)$; these $c+1$ multinomial distributions are independent. Up to an additive constant, the log-likelihood is:

$$\sum_i \sum_j n_{ij}\left(\ln\left(p_{ij}\right) - \ln\left(p_j\right)\right) + \sum_i m_i \ln\left(\pi_i\right)$$

Under the usual constraints on the set of probabilities p_{ij} it can be shown that this function has a unique maximum, either inside the parameter space or on its boundary. In both cases, this maximum likelihood solution can be obtained by a suitable algorithm, for instance by the constrOptim procedure in the R package (R Development Core Team, 2008). We shall call this estimating method "Maximum Likelihood 2" (ML2).

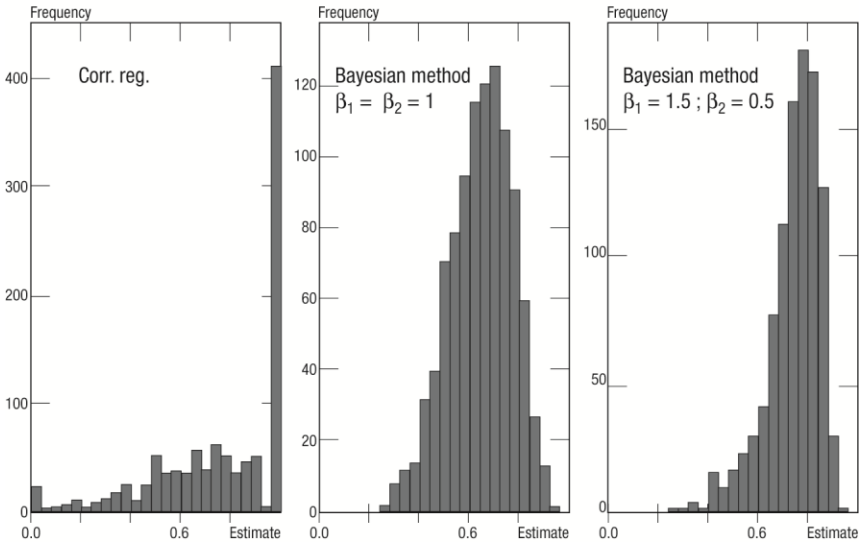
In practice, the parameters of interest are the p_j $(j = 1,\ldots,c)$. It is worth noting that, even if the maximum of the likelihood is reached on the boundary (i.e. at least one of the p_{ij} is equal to zero), this does not mean that the corresponding p_j vanishes since this is only the case if $p_{ij} = 0$ *for all* $i = 1, \ldots, l$.

For the Bayesian model, we first take "neutral" prior parameters $\beta_1 = \beta_2 = 1$. We obtain estimates with mean 0.64, standard deviation 0.12, and mean squared error 0.05. The corresponding histogram is given in Figure 69 (centre).

It can be seen that the Bayesian method is more satisfactory, even when the situation is highly unfavourable for it, with a probability to be estimated relatively close to unity and a prior distribution that allocates a mean of 0.5.

If it is known in advance that the probability to be estimated is "fairly high", this may be allowed for in the value of β_1; to check the impact we repeat the estimation with $\beta_1 = 1.5$ and $\beta_2 = 0.5$. We obtain a mean estimate of 0.78, standard deviation 0.11 and mean squared error 0.01. The estimates are consequently much better, and this is also illustrated in Figure 69 (right). In general terms, therefore, if one has some idea of the age distribution of the observed population , it should be introduced into the model without hesitation. However, in practice, it must be borne in mind that the choice of the prior distribution must be justified.

Figure 13.1. Simulations – Example 1: histograms of estimates of the probability of belonging to the first age group obtained in 1,000 iterations. Left: "corrected" regression; centre: Bayesian method with $\beta_1 = \beta_2 = 1$; right: Bayesian method with $\beta_1 = 1.5$ and $\beta_2 = 0.5$



Example 2

This example corresponds very closely to the second theoretical example (3 stages and 2 ages) in the previous chapter. The conditional probabilities of the three stages are (0.6250 0.3125 0.0625) for age class 1 and (0.125 0.375 0.500) for age class 2; the marginal probabilities for the three stages are (0.500; 0.328; 0.172) and the marginal probabilities for the ages are 0.755 and 0.245. These last two probabilities are to be estimated from multinomial samples of stages of size t (chosen as 20). The estimated results are given for the first probability (here p), whose true value is 0.755.

Since in this case l and c are different, the IALK method and the corrected regression do not necessarily give the same results, so it is instructive to compare them. In this comparison, we also introduce the Maximum Likelihood 2 method (Box 10) and our Bayesian method with $\beta_1 = \beta_2 = 1$. The histograms of the results are shown in Figure 70 and the key features in Table 58.

Figure 13.2. Simulations – Example 2: histograms of estimates obtained by four methods (with $\beta_1 = \beta_2 = 1$ for the Bayesian method)

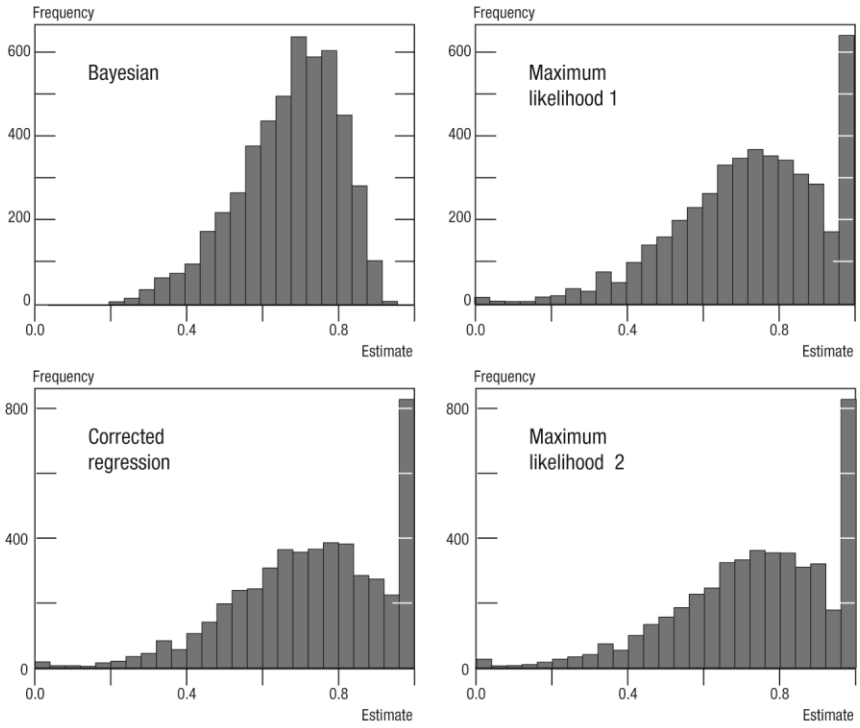


Table 13.1 Simulations – Example 2 – 5,000 iterations. Characteristics of the estimates of *p* obtained by four different methods

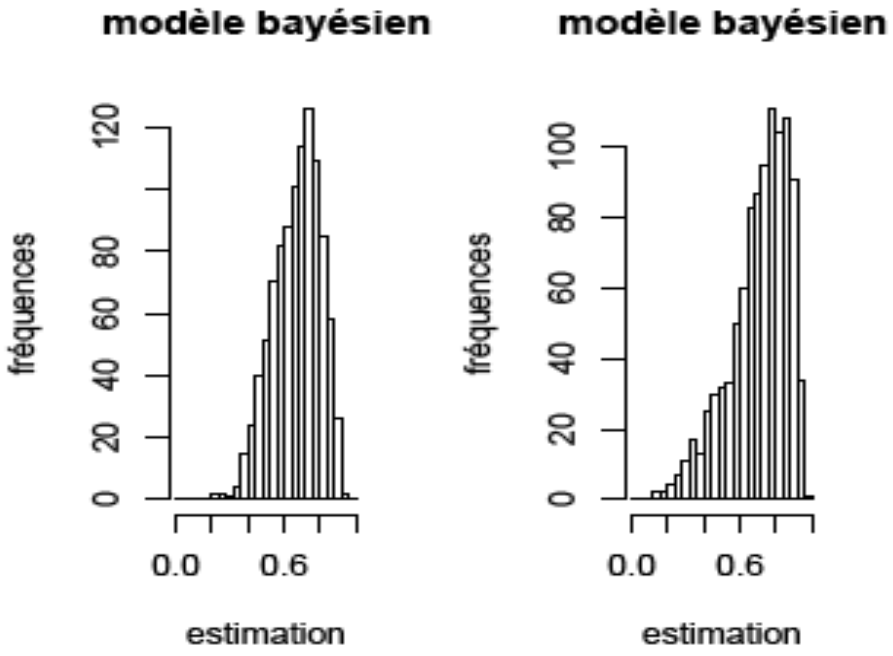
	Mean	Standard deviation	MSE
Bayesian method	0.672	0.134	0.025
Corrected regression	0.724	0.201	0.041
Maximum Likelihood 1	0.727	0.197	0.039
Maximum Likelihood 2	0.737	0.205	0.042

As we said above, we use the term "Maximum Likelihood 1" for what is more generally known as the IALK method. This is because the name IALK confuses the concept with a numerical solution technique which is incomplete in any case. The I in IALK basically refers to an "iterative" process that only gives a clear result if the maximum likelihood lies within the set of possible solutions and, in that case, corresponds to the zero point on the likelihood gradient; we prefer to take the principle of the maximum likelihood method to its logical conclusion and also consider a maximum at the boundary. Although the IALK iterative process probably provides the maximum, this has not, to our knowledge, been rigorously proven, so we look for the maximum likelihood *in all cases* by using the constrOptim procedure from the R package (R Development Core Team, 2008). The same procedure was used to find the maximum likelihood in the more general model underlying Maximum Likelihood 2.

Table 58 shows that the mean squared errors are similar for the corrected regression and the two maximum likelihood methods, while that of the Bayesian method is significantly lower, although the prior distribution uses no indication as to the true value of *p*. This in fact explains why this method gives the largest bias, fortunately corrected by a much lower variance.

Maximum Likelihood 1 gives slightly better results than corrected regression, which is to be expected, since it accounts better for the nature of the sampling errors. However, it may at first glance seem surprising that Maximum Likelihood 2 is not better than ML1 since, here again, it accounts better for the nature of the errors; the reason is probably that the number of parameters to be estimated is less parsimonious, which does reduce the bias but increases the variance more, and thus increases the mean squared error. Note finally that the individual differences between the estimates of the last three methods are quite small, rarely more than 0.1; in this particular example, regression and Maximum Likelihood 1 relatively often provide estimates of *p* "at the boundary", with unity in more than 10% of cases and even a certain number of zero values (approximately 0.2%). Maximum Likelihood 2 has the advantage of leading less often to these results (in our simulations we obtained no zero estimates and only 1.6% at unity).

Figure 13.3. Simulations – Example 2. Histograms of estimates obtained for *p* by the Bayesian method with $\beta_1 = \beta_2 = 1$ (left) and $\beta_1 = \beta_2 = 0.5$ (right).



Up to now we have considered the Bayesian method with $\beta_1 = \beta_2 = 1$ and have seen that it performs better than methods based on other principles. However, we still need to examine the influence of the parameters of the prior distribution. To avoid giving undue advantage to the method we are comparing, we have taken various values for these prior parameters, staying within "neutral" prior means (0.50 0.50) but varying the confidence levels, with successively 1, 0.75 and 0.50 as common values for β_1 and β_2. A higher value for β_j (1.2) was also envisaged.

The characteristics of the distribution of estimates obtained for *p* are given in Table 13.2.

Table 13.2. Simulations – Example 2. Characteristics of *p* estimates by the Bayesian method for prior values of β_1 and β_2.

	Mean	Standard deviation	MSE
$\beta_1 = \beta_2 = 1.2$	0.660	0.123	0.024
$\beta_1 = \beta_2 = 1$	0.672	0.129	0.024
$\beta_1 = \beta_2 = 0.75$	0.688	0.149	0.027
$\beta_1 = \beta_2 = 0.50$	0.707	0.163	0.029

Figure 13.3 also compares the histograms obtained with $\beta_1 = \beta_2 = 1$ (left) and $\beta_1 = \beta_2 = 0.5$ (right).

In both the numerical values in Table 13.2 and the histograms in Figure 13.3, it can be seen that, just as with the previous example, reducing the β_j slightly reduces the bias (which is due to a mean choice of prior probability lower than the true value), but at the cost of a noticeable increase in variance; in all, this ultimately gives a slight deterioration in mean squared error (MSE). Increasing the β_j above unity increases the bias and gives an equivalent mean squared error. There appear to be no strong arguments for any particular choice of β_j, but the simple option $\beta_1 = \beta_2 = 1$ can no doubt be recommended with no great risk.

Example 3

Now we apply simulation to an example that comes closer to the problems encountered in paleodemographic practice. We began with the 7 × 7 example of the Maubuisson nuns with conditional reference probabilities deduced from the frequencies given in Table 56 of the previous chapter, and row (stage) probabilities of

(0.180 0.068 0.115 0.159 0.119 0.188 0.171).

These probabilities comply with probabilities of dying in each age class of

(0.012 0.025 0.087 0.170 0.289 0.210 0.207)

as calculated from an exhaustive evaluation of deaths in the period 1670-1789 recorded in the registers available.

The simulation results are therefore to be compared with the second set. The simulation involves $R = 1{,}000$ iterations, the multinomial samples comprise 37 individuals for the site data and are the same size as the reference data samples for the latter.

We first compare the estimates obtained by the Bayesian method (posterior means) with $\beta = (1, 1, 1, 1, 1, 1, 1)$, by regression (with a positivity constraint), and by Maximum Likelihood 1 and 2. Table 13.3 shows the mean of estimates for each age class and each method, and, for each method, the total over seven age classes of mean squared errors (total MSE) and this same total weighted by the true values of the probabilities (relative MSE).

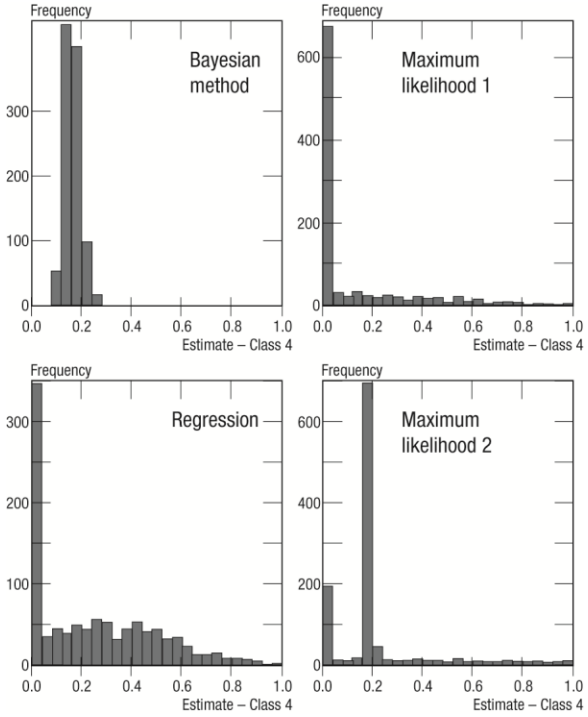
Table 13.3. Simulations – Example 3. Means of probability estimates for each age class obtained by four methods, and total and relative MSEs.

Method	Age class							Total MSE	Relative MSE
	20-29	30-39	40-49	50-59	60-69	70-79	80+		
Bayes, uniform prior	0.078	0.098	0.102	0.143	0.194	0.182	0.203	0.033	0.751
Regression	0.001	0.069	0.061	0. 232	0.032	0.402	0. 203	0.233	1.526
Max. L. 1	0.035	0.056	0.063	0.137	0.253	0.232	0.222	0.253	2.204
Max. L. 2	0.174	0.071	0. 155	0. 257	0.168	0.125	0. 051	0.170	3.409
True values	0.012	0.025	0.087	0.170	0.289	0.210	0.207		

To round out the raw numerical data above, Figure 13.4 gives the histograms of the frequencies of estimates obtained by the four methods for the probability of a single age-class, class 4 ("true" value 0.17).

It can be seen that the Bayesian method clearly outranks the other three in terms of mean squared error (whether total or relative). For the other three methods, theory predicts that where $l = c$, regression with positivity constraint and Maximum Likelihood 1 should provide equivalent results, at least when the estimates are not at the boundary of admissible values. And this is indeed observed with other examples not described here. In this case, however, the two methods provide fairly dissimilar results; most likely because the results are all at the boundary (one estimated probability is zero). It is also clear that the optimisation of the functions concerned is highly unstable, which is a further argument against these methods. With respect to zero estimates, Figure 72 shows that there are a fair number of them with regression and Maximum Likelihood 1, even for Class 4, where the value to be estimated (0.17) is not particularly low.

Figure 13.4. Simulations – Example 3. Histograms of estimates obtained from 1,000 iterations by four methods for probability p_4 (with uniform prior for the Bayesian method).



Although the Bayesian method is clearly superior to the others, it presents notable biases, particularly in estimating the low probabilities of classes 1 and 2. The method could be improved in two ways. One is to adjust the "weighting" of the prior distribution by modifying the variance while keeping the same means (equal for each class). It is clear, however, that the observed biases are produced by a prior distribution highly unfavourable in its means since it allocates the same probability of death to each age class. It would be more realistic to allocate to each class a prior mean equivalent to standard mortality rates, and the data collected at a given site would serve to modify the standard for that site.

To test this first attempted improvement, we adjust the variances of the prior distribution without changing the means; the β_j remains the same for all j but takes successive values 1.25 , 1, 0.75 and 0.50, so that the variances gradually increase. The means of the estimates obtained for these four prior distributions with 1,000 iterations are given in Table 61, with total and relative mean squared errors in the last two columns as before.

It can be seen that $\beta_j = 1$ is a reasonable compromise. Low probabilities tend to be overestimated and high probabilities underestimated; the overestimation of the low probabilities is less marked when the β_j are smaller, giving a lower relative mean squared error, but what is gained in one place is lost in another, so the total mean squared error is higher.

Table 13.4. Simulations – Example 3. Mean of the estimates by the posterior mean from identical β_j varying from 0.50 to 1.25, and total and relative mean squared errors.

Value of β_j	Age class							Total MSE	Relative MSE
	20-29	30-39	40-49	50-59	60-69	70-79	80+		
1.25 for all *j*	0.082	0.102	0.105	0.143	0.190	0.177	0.200	0.030	0.811
1 for all *j*	0.078	0.098	0.102	0.143	0.194	0.182	0.203	0.033	0.751
0.75 for all *j*	0.0678	0.087	0.093	0.141	0.205	0.190	0.216	0.036	0.672
0.5 for all *j*	0.059	0.079	0.085	0.141	0.215	0.200	0.221	0.042	0.598
True probabilities	0.012	0.025	0.087	0.170	0.289	0.210	0.207		

We now attempt to use standard mortality rates instead of equal prior means. In view of the nature of this example, we take the pre-industrial standard mortality for women, which, with considerable rounding, gives the following proportions for the seven classes:

0.10 0.11 0.12 0.15 0.21 0.21 0.10

The β_j are calculated in proportion to these values, with a sum $\beta_.$ varying around 7 as suggested by the previous study.

The results obtained for $\beta_.$, successively equal to 5, 7 and 10, are given in Table 62, with comparative figures for a uniform prior distribution.

The first point to note is that the new prior results in a substantial improvement, in both absolute and relative terms, particularly in relative errors because of the major bias in estimating low probabilities with a uniform prior distribution. Comparison of the three prior distributions deduced from the standard shows relatively similar behaviour, with an advantage for a smaller $\beta_.$ if focusing on relative errors, for a larger $\beta_.$ if focusing on absolute ones. The choice of $\beta_. = 7$ (number of columns) appears to be a good compromise and our conclusion is to recommend it.

Table 13.5. Simulations – Example 3. Comparison of mean estimates obtained from a uniform prior distribution (row 1) and three prior distributions deduced from the female pre-industrial standard, with total and relative mean squared errors.

Value of β_j	Age class							Total MSE	Relative MSE
	20-29	30-39	40-49	50-59	60-69	70-79	80+		
1 for all *j*	0.078	0.098	0. 102	0.143	0.194	0.182	0.203	0.033	0.751
Standard $\beta_. = 5$	0.054	0.074	0.084	0.149	0.260	0.239	0.140	0.028	0.428
Standard $\beta_. = 7$	0.059	0.079	0.089	0.148	0.254	0.235	0.135	0.024	0.451
Standard $\beta_. = 10$	0.065	0.084	0.094	0.148	0.246	0.233	0.130	0.022	0.498
True probabilities	0.012	0.025	0.087	0.170	0.289	0.210	0.207		

Actually, for the Maubuisson nuns whose ages at death are simulated in this example, the fact that the site is a convent cemetery provides important supplementary information because the young nuns were probably in better health on average than the general population and not exposed to certain major mortality risks, particularly death in childbirth. The method can incorporate this prior information by modifying the β_j parameters. For example, we may consider that mortality in the 20-29 age class is probably more than halved and that mortality in the following age class is also halved. This leads us to consider a new β_j vector $(\frac{7}{6.21})$ (0.30 0.40 0.84 1.05 1.47 1.47 0.70) (the coefficient $\frac{7}{6.21}$ is used to bring total $\beta_.$ to 7 as recommended above). This produces the following mean estimates:

0.033 0.051 0.107 0.164 0.262 0.245 0.138

with a total mean squared error of 0.019 and a relative mean squared error of 0.183.

The improvement is significant, most clearly for the low probabilities in Classes 1 and 2, which are most affected by the change in prior distribution, and consequently for the relative mean squared error.

Finally, it is reasonable in this case to use the Bocquet-Appel and Bacro method (2008) with the ProbAtri20-90 set of 756 base vectors. Table 13.6 gives the total and relative mean squared errors for the results obtained with that method and the Bayesian method using the β_j above (MPI means "modified" pre-industrial); for a comprehensive comparison we also give the results from Table 3. 8 for the β_j corresponding to the pre-industrial (PI) standard.

Comparing our method and that of Bocquet-Appel and Bacro on this example, it can be seen that their performances in this case are similar, with a slight advantage to the Bayesian method if an informed choice of prior distribution is possible.

Table 13.6. Simulations – Example 3. Comparison of the Bayesian method using two prior distributions (PI: female pre-industrial standard; MPI: modified pre-industrial standard, see text) with the Bocquet-Appel and Bacro method.

Method	Total MSE	Relative MSE
Bayes (PI)	0.024	0.451
Bayes (MPI)	0.019	0.183
Bocquet-Appel	0.021	0.304

Example 4

Here we take an example where the bone stages are subdivided into 5 categories, with the same 7 age classes as before. Traditional frequentist methods cannot be used because there are more columns than rows, but it is possible to use the Bocquet-Appel and Bacro method with the ProbAtri20-90 set of 756 vectors. This example serves to continue the comparison between that method and ours.

The reference data table is the 5 × 7 table for both sexes as follows:

138 68.8 58.2 33.2 13.4 7.0 5.0

42 58.6 54.6 48.8 24.2 26.4 14.6

18 25.4 35.4 38.2 31.0 33.8 24.0

12 17.0 20.4 35.6 49.2 42.0 36.6

4 16.4 12.4 24.6 42.2 32.4 42.8

We took in turn three different probability vectors for the age classes

(0.166 0.115 0.150 0.178 0.173 0.134 0.084)

(0.10 0.10 0.15 0.15 0.20 0.20 0.10)

(0.35 0.09 0.09 0.10 0.18 0.11 0.08)

They were chosen to represent realistic situations, assumed to be favourable to one method or the other. The first is the vector mean of the ProbAtri20-90 set of 756 vectors; the second a vector close to the pre-industrial standard; and the third the estimate made for the Frénouville site. We consider that no particular prior information is available: the Bayesian method is therefore used with a prior distribution complying with the pre-industrial standard and $\beta_{\cdot} = 7$.

Table 13.7 gives the squared differences observed from a simulation of 1,000 repetitions. It can be seen that the two methods perform more or less equally for the last case, while ours is clearly better for the other two (including the first case where intuition might have suggested otherwise).

Table 13.7. Simulations – Example 4. Comparison of the Bayesian method (with pre-industrial standard prior) and the Bocquet-Appel and Bacro method.

	Case 1		Case 2		Case 3	
	Total MSE	Relative MSE	Total MSE	Relative MSE	Total MSE	Relative MSE
Bayes (PI)	0.001	0.006	0.004	0.028	0.035	0.231
Bocquet-Appel	0.020	0.155	0.022	0.182	0.034	0.248

Conclusion to the simulated examples

The examples provide initial data for a discussion of the practical aspects of choosing the prior distribution. They also show that our method is clearly preferable to any method that does not address the specific features of the problem posed. Compared with the Bocquet-Appel and Bacro method, which makes wide use of these features, our method appears on the whole to be perfectly competitive, and much simpler to apply: the choice of a prior distribution is clearly easier and more flexible than constructing a set of base vectors. However, this comparison merits further examination (on this point, see Caussinus and Courgeau, 2010).

Note that the comparisons above concern the effectiveness of the method in producing point estimates. In the section below, we demonstrate a few more of its advantages.

13.4. Examples of archaeological application

We now apply our Bayesian method to the two archaeological examples addressed differently in the previous chapter. The choice of the prior distribution will use the principles described in Section 13.3. And a further question will be addressed: how to weight the reference data. This was not relevant in the above simulations because the invariance model was assumed to be valid by definition.

Example 1: Loisy-en-Brie population

The data are those considered in the previous chapter, for which regression was used (or, equivalent in this case, IALK). There are six age classes of equal duration and six stages. If we have no precise prior information, we can first apply the Bayesian method with $\beta = (1, 1, 1, 1, 1, 1)$. Table 13.8 gives the estimated proportions for each class obtained for two values of coefficient r (weighting of reference data): 1 and 0.75. The standard deviations of the posterior distributions are also given.

Table 13.8 Loisy-en-Brie example

Estimated parameters ${}_{12}p_{23},\ldots,{}_{12}p_{83}$ with $\beta_1 = \beta_2 = \ldots = \beta_6 = 1$

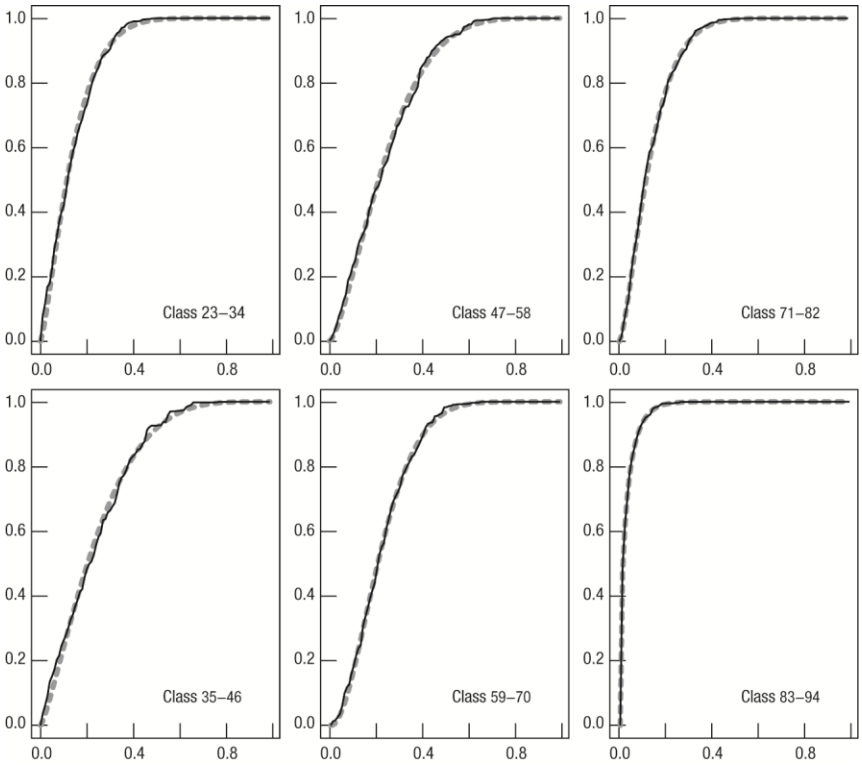
	Age class	23-34	35-46	47-58	59-70	71-82	83-94
$r = 1$	Expected posterior probability	0.148	0.252	0.245	0.174	0.114	0.067
	Standard deviation	0.099	0.162	0.171	0.122	0.094	0.059
$r = 0.75$	Expected posterior probability	0.153	0.247	0.238	0.175	0.116	0.071
	Standard deviation	0.101	0.159	0.165	0.122	0.097	0.062

.

The differences in expected posterior probabilities according to r are negligible if one takes account of the standard deviations (note that the standard deviations themselves differ little). There seems no reason, therefore, to weight the reference data and we continue the study with $r = 1$.

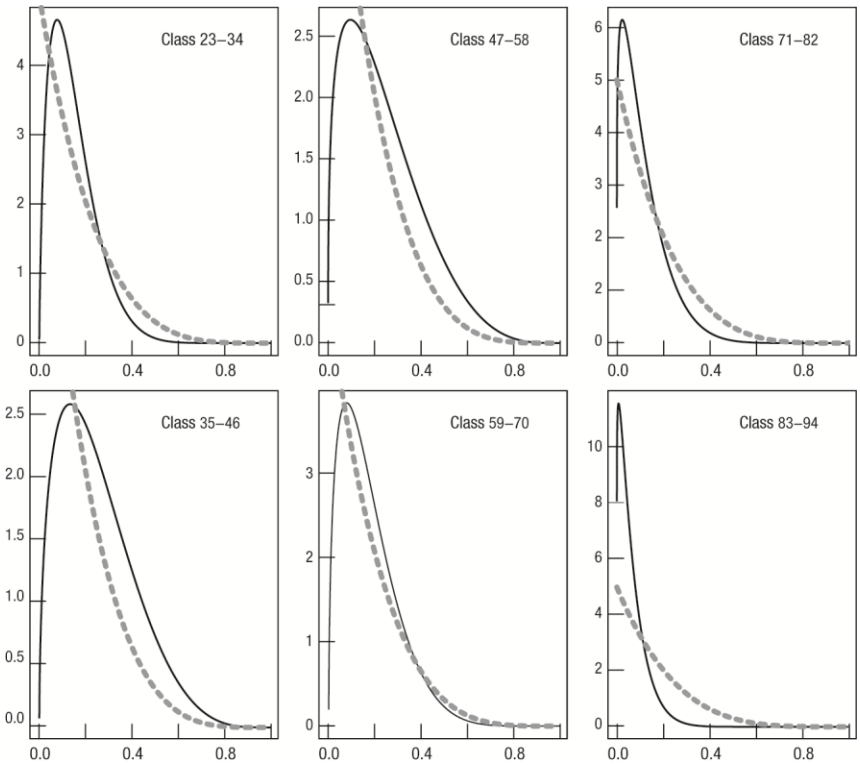
From the posterior means and standard deviations we can establish the Beta distributions approximating the posterior distribution of each of the 6 probabilities of belonging to each age class (see Sect. 13.2.1). We first examine how satisfactory this approximation was by also calculating the exact distribution functions by the method given in 2.1. Figure 13.5 compares the exact and approximate distribution functions. It can be seen that the approximation is extremely close, as is also suggested by the relative difference between the 3rd-order (to the power 1/3) and 4th-order (to the power ¼) moments: they are all less than 1.8% for the 3rd order and less than 3.8% for the 4th order. In practice the Beta distribution approximation appears mainly to provide an opportune sort of smoothing.

Figure 13.5. Loisy-en-Brie example (6 age classes). Exact (solid line) and Beta approximate (dotted line) posterior distribution functions.



In the light of these results, we consider the Beta approximations of the posterior distributions. The corresponding densities are shown in Figure 13.6, compared with the prior densities (all Beta (1; 5) densities, mean 0.167 and standard deviation 0.141).

Figure 13.6. Loisy-en-Brie example. Posterior probability density (solid line) for each age class, compared with prior density (dashed line), in the case of uniform prior distribution.



We shall not enter here into a detailed discussion of these initial results because the prior distribution considered here does not integrate the fact that we are attempting to estimate a mortality distribution. Since we have no other specific *a priori* argument, it is reasonable to fit the prior distribution on the pre-industrial standard. We did this by assuming that parameters β_j are proportional to that standard and sum to 6 (see Section 13.3) as follows:

$\beta = (0.77 \;\; 0.90 \;\; 1.16 \;\; 1.53 \;\; 1.25 \;\; 0.39)$

The posterior means and standard deviations are now the ones shown in Table 13.9.

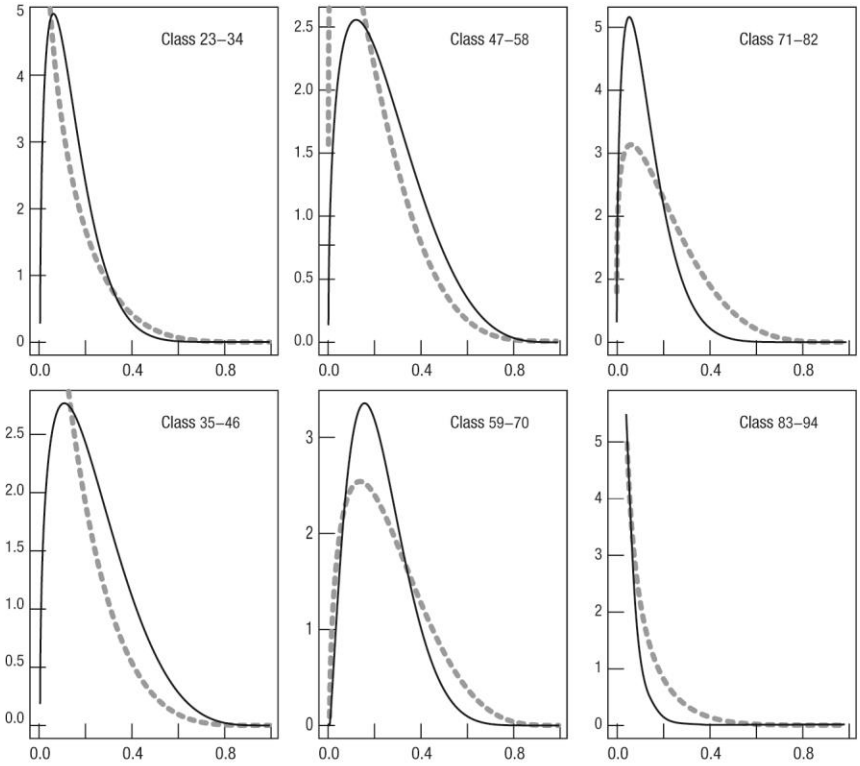
Table 13.9. Loisy-en-Brie example.

Estimated parameters ${}_{12}p_{23},\ldots,{}_{12}p_{83}$ with prior deduced from pre-industrial standard.

Age class	23-34	35-46	47-58	59-70	71-82	83-94
Expected posterior probability	0.139	0.233	0.250	0.220	0.132	0.026
Standard deviation	0.100	0.157	0.165	0.123	0.096	0.038

As before, the Beta approximation of the posterior distributions is excellent. We used it therefore for the densities shown in Figure 13.7.

Figure13.7. Loisy-en-Brie example. Posterior probability density (solid line) for each age class, compared with prior density (dashed line), in the case of a prior complying with the pre-industrial standard.



Comparing the numerical values in Tables 13.8 and 13.9 as well as the posterior densities in Figures 13.6 and 13.7, most of the estimated probabilities are quite stable, demonstrating the limited influence of this prior on the result of the estimation. The greatest difference is observed for the 83-94 age class, which is to be expected since this class will clearly have a low probability, a feature allowed for by the second prior distribution but not the uniform one. With the uniform prior, the posterior deviates quite significantly from it, showing that the data impose a serious downward revision; the same is true, though to a lesser extent, for the second prior distribution (taking only the mean, it falls from 0.39/6 = 0.065 to 0.026), confirming that the corresponding probability is not only low but in all likelihood lower than the pre-industrial standard. The next greatest difference between the two estimates can be seen in the 59-70 age class, where the posterior mean rises with the pre-industrial standard, as does the prior mean; the posterior standard deviation also increases (in fact the posterior density is quite clearly more "open"). But these differences are limited, if we consider the wide dispersion of the posterior distributions, due to the small sample size and the structural instability of the problem considered.

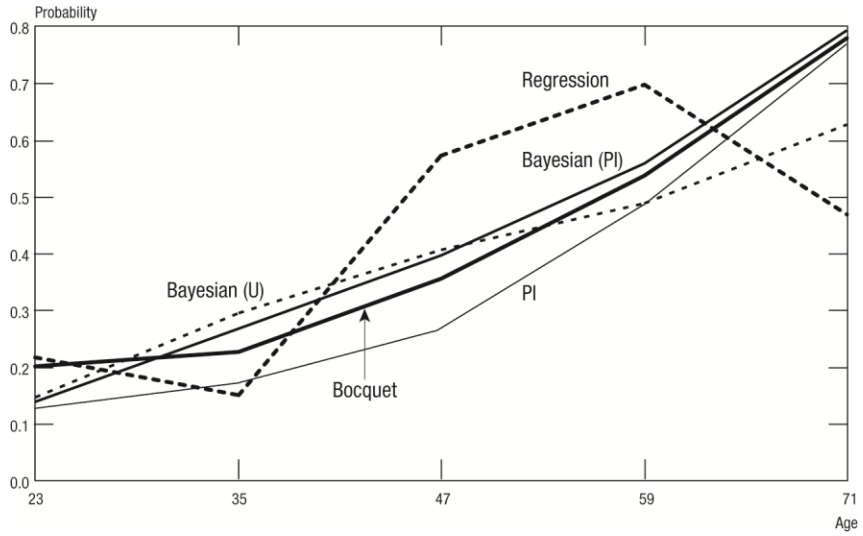
For a clearer idea of the accuracy of the estimates, credibility intervals can be calculated. As we have already pointed out, because of the considerable asymmetry of the distributions involved, it is highly inadvisable to calculate symmetric confidence intervals of the "mean plus or minus so many standard deviations" type. It is better to stay with the Bayesian paradigm and give quantiles of the posterior distribution. Table 13.10 gives quantiles 0.05 and 0.95, which provide a 90% credibility interval and quantiles 0.25 and 0.75 (quartiles), which provide a 50% credibility interval.

Table13.10. Loisy-en-Brie example. Estimates by posterior mean and quantiles for 90% and 50% credibility intervals

Estimated probabilities	${}_{12}p_{23}$	${}_{12}p_{35}$	${}_{12}p_{47}$	${}_{12}p_{59}$	${}_{12}p_{71}$	${}_{12}p_{83}$
Posterior mean	0.139	0.233	0.250	0.220	0.132	0.026
Quantile 0.05	0.019	0.031	0.035	0.053	0.018	0.000
Quantile 0.95	0.334	0.536	0.566	0.452	0.319	0.104
Quantile 0.25	0.062	0.108	0.119	0.126	0.059	0.002
Quantile 0.75	0.193	0.329	0.353	0.296	0.187	0.035

Now we shall compare these results with those obtained by the various methods presented in the previous chapter. Rather than work with deaths observed for various age groups, it is useful to observe a quantity more generally used in demography: the probabilities of death per age group, which can be estimated from mortality data on the assumption of a stationary population. We examine the various estimates made for the Loisy-en-Brie population. Figure 13.8 gives these various probabilities.

Figure13.8. Probabilities of death for Loisy-en-Brie estimated by the Bayesian method with pre-industrial standard prior (Bayesian PI), uniform prior (Bayesian U), regression, and the method proposed by Bocquet-Appel and Bacro in 2008 (Bocquet), compared with the pre-industrial standard (PI).



The figure clearly shows that the curve of regression method estimates, which, it will be recalled, gives the same results as the IALK method, is highly erratic. This confirms what is seen in the simulations, where the regression method leads to widely dispersed results (see Section 3). On the other hand, the Bayesian methods, with either a uniform or a pre-industrial standard prior, exhibit a regular increase in the probabilities of death with age. The difference in prior distribution only affects the last two age groups, and then only slightly. We enter the probabilities of death for this standard, clearly showing its effect on the Bayesian estimates

for the oldest age group. Despite this slight effect, the Bayesian estimated distributions differ significantly from the standard and are similar to each other, showing the robustness of this method, whatever the prior distribution chosen. We explained above why we prefer the second estimate (PI). We also entered the probabilities of death obtained by the method proposed in Bocquet-Appel and Bacro (2008: see presentation and results in the previous chapter): their distribution is fairly close to the Bayesian solution with the pre-industrial standard. But it is higher for the youngest age group and lower for the others. On this precise point, the fact that our estimates are stable with respect to the prior distribution and that the dispersion of the posterior distribution is relatively low suggests that they can be ascribed to a high confidence level. As for the overall results, it can be seen that the Bocquet-Appel and Bacro method with the choice of a parametric space comprising a distribution of age distributions close to the Bayesian solution, leads to results similar to the Bayesian method, but at the cost of a significantly more cumbersome technique.

Example 2: the Maubuisson nuns (17th-18th centuries)

We now turn again to the Maubuisson example, which gave unacceptable results with the ordinary least squares method, such as negative values and values above unity. Even if the least squares can be forced to meet the necessary constraints, the estimates obtained are on the boundary of the parametric space (zero values) and clearly unrealistic. We adopt now the same division into 7 stages and 7 age classes as before.[4] The numbers observed for the various stages in a sample of 37 skulls are (6 2 4 5 3 9 8).

We have a large amount of prior information about this site, particularly useful since the sample is fairly small (37). These are nuns, and consequently all women, all theoretically older than 20. We can therefore opt for specific reference data, namely those already used in the previous chapter and Section 13.3 of this chapter. If we stick to this information, using the experience from Section 13.3, Example 3, we shall take for the 7 parameters to be estimated a Dirichlet prior probability distribution with β_j parameters proportional to the values of the pre-industrial standard (women) and summing to 7, namely (0.70 0.77 0.84 1.05 1.47 1.47 0.70). We present here these initial estimates, more for comparative purposes than for a conclusion. The fact that these women were nuns gives us further information: on admission they were for many reasons in better health than the mean of the general population; they were then protected from various major mortality risks, particularly death in childbirth. These factors can be considered to reduce the mortality of the 20-29 age class by just over 50% and of the 30-39 class by just under 50%, thus replacing the parameters of the prior distribution by (0.30 0.40 0.84 1.05 1.47 1.47 0.70) or rather by the proportional values (0.337 0.449 0.944 1.180 1.652 1.652 0.786) summing to 7, as recommended in Section 13.3. From this prior distribution we propose a second estimation: it appears prima facie to be the one that should be adopted in practice, and we shall see how far the results obtained confirm this.

Finally, as mentioned in the previous chapter, there is a further major source of information in this case: the convent records give direct evidence of the actual ages at death; as a result the age class probabilities may be evaluated as follows: (0.012 0.025 0.087 0.170 0.289 0.210 0.207). We have therefore an objective way of judging the effectiveness of the method,

[4] A study of this site with 5 suture stages instead of 7 and five-year age classes (a total of 13) has been carried out and is given in Séguy et al. (2012). The results tally completely, showing in particular that the method described here works effectively with significantly more age classes than stages.

although some caution is necessary, because the evaluation is probably only approximate and the 37 skulls are only a small and possibly biased sample.[5]

We begin with an analysis using a prior distribution that complies with the pre-industrial standard for women. We applied various "reduction coefficients" to the reference data; since the results were observed to be stable, we shall stick with coefficient 1. The posterior expected values and standard deviations are given in Table 13.11.

Table13.11. Maubuisson example. Estimated parameters ${}_{10}p_{20}$, to p_{80+}: posterior means and standard deviations for a "standard" prior distribution.

	${}_{10}p_{20}$	${}_{10}p_{30}$	${}_{10}p_{40}$	${}_{10}p_{50}$	${}_{10}p_{60}$	${}_{10}p_{70}$	p_{80+}
Posterior expected value	0.048	0.067	0.071	0.135	0.301	0.219	0.159
Posterior standard deviation	0.050	0.068	0.069	0.114	0.166	0.142	0.135

It is instructive to compare the posterior means with the prior means, in this case (0.10 0.11 0.12 0.15 0.21 0.21 0.10). It can be seen that the data significantly revise downwards the probabilities for the "youngest" classes, and upwards only the two oldest ones; which is consistent with the discussion above. We leave this analysis as it stands and move on to the Bayesian analysis with a prior distribution with parameters (0.337 0.449 0.944 1.180 1.652 1.652 0.786) corresponding to a modified pre-industrial standard as discussed above. The posterior means and standard deviations obtained are given in Table 13.12. We now go on to examine various other parameters of the posterior distribution relating to each age class. For example, we may calculate quantiles: a selection is given in Table 13.13. Figure 13.9 is a graphical representation of the 50% credibility intervals, comparing posterior means and medians with the values in the records.

[5] Note, however, that with the reference probabilities we are using, the sample is fully compatible with the documented values. If we calculate theoretical frequencies for stages from these data and compare them with the observed values by chi-squared test, we obtain 1.93 with 6 degrees of freedom.

Table13.12. Maubuisson example. Estimated parameters ${}_{10}p_{20}$, to p_{80+}: posterior means and standard deviations for a "modified pre-industrial standard" prior distribution

	${}_{10}p_{20}$	${}_{10}p_{30}$	${}_{10}p_{40}$	${}_{10}p_{50}$	${}_{10}p_{60}$	${}_{10}p_{70}$	p_{80+}
Posterior expected value	0.025	0.041	0.083	0.151	0.311	0.230	0.159
Posterior standard deviation	0.037	0.054	0.074	0.119	0.163	0.142	0.132

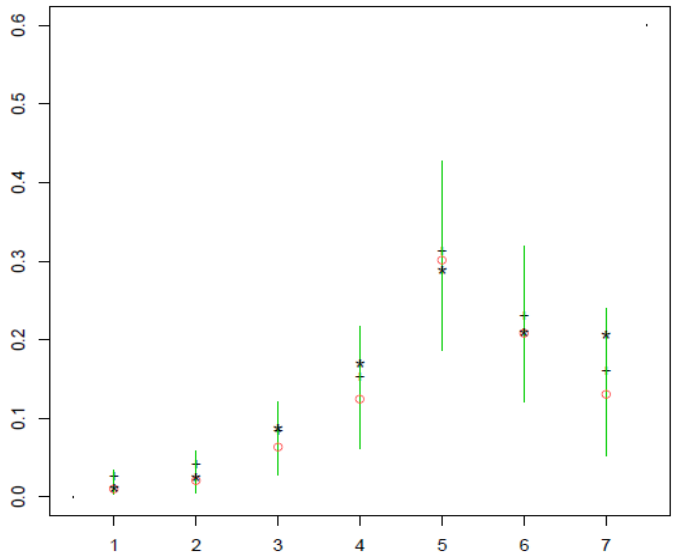
.

Table13.13. Maubuisson example. Posterior medians and quantiles for 90% and 50% credibility intervals

Estimated probability	${}_{10}p_{20}$	${}_{10}p_{30}$	${}_{10}p_{40}$	${}_{10}p_{50}$	${}_{10}p_{60}$	${}_{10}p_{70}$	p_{80+}
Median	0.010	0.020	0.065	0.125	0.300	0.208	0.130
Quantile 0.05	0.002	0.002	0.006	0.015	0.069	0.042	0.008
Quantile 0.95	0.106	0.156	0.232	0.387	0.604	0.501	0.417
Quantile 0.25	0.003	0.005	0.028	0.061	0.186	0.121	0.053
Quantile 0.75	0.034	0.058	0.121	0.216	0.427	0.319	0.239

.

Figure13.9. Maubuisson example. Fifty-percent credibility intervals for 7 age classes (green lines), posterior means (black dashes) and medians (red circles) compared with values from the records (asterisks).



The prior means in this case are

(0.048 0.064 0.135 0.169 0.236 0.236 0.112)

and the target values given in the records are

(0.012 0.025 0.087 0.170 0.289 0.210 0.207).

First, it is clear that with so small a sample it is not possible to obtain very precise estimates, as can be seen from the rather wide credibility intervals. But some relevant information may be inferred from the analysis of the data. It can be seen that this analysis leads to a further downward revision of probabilities for the first three classes and a further upward revision for the fifth and seventh. The most noticeable differences (particularly in relative terms) between posterior means and target values are to be seen for the first two probability figures, where the values taken from the records are much lower than would have been expected. In fact, with the highly asymmetric distributions corresponding to extremely low probabilities, the mean can be deceptive; if we examine class 1, for example, we can see that 50% of the posterior probability lies in the interval [0.003 0.034], whose mean 0.0185 is close to the target value, as is, indeed even more so, the posterior median 0.010; the same holds for class 2. It is therefore justifiable to suppose that for these two classes the posterior means overestimate the true values, an indication that turns out to be realistic. In more general terms, it can be seen that the 50% credibility intervals do indeed cover the target values, which are always close to the posterior mean and median, and the greatest discrepancy, albeit quite understandable given the sample size, is to be found in class 7. To sum up, we may say that the posterior means provide base information likely to be usefully supplemented by a range of considerations concerning various features of posterior distributions (medians, credibility intervals, etc.).

Here our results can be compared with those obtained by Bocquet-Appel and Bacro's Iterage algorithm, because the division into classes adopted corresponds to the ProbAtri20-90 file of "prior" vectors they provide. Their algorithm gives the following estimates:

0.025 0.036 0.073 0.133 0.209 0.268 0.255

It can immediately be seen that the estimates for the first two classes are closer to the target values (probably because we did not wish to give too low a prior mean for these classes), but that the estimates for the other classes are on the whole better with our method. To have a closer idea, we calculated the distances between estimates and target values in two ways: sum of squared deviations and sum of squared deviations weighted by the target value. The figures are

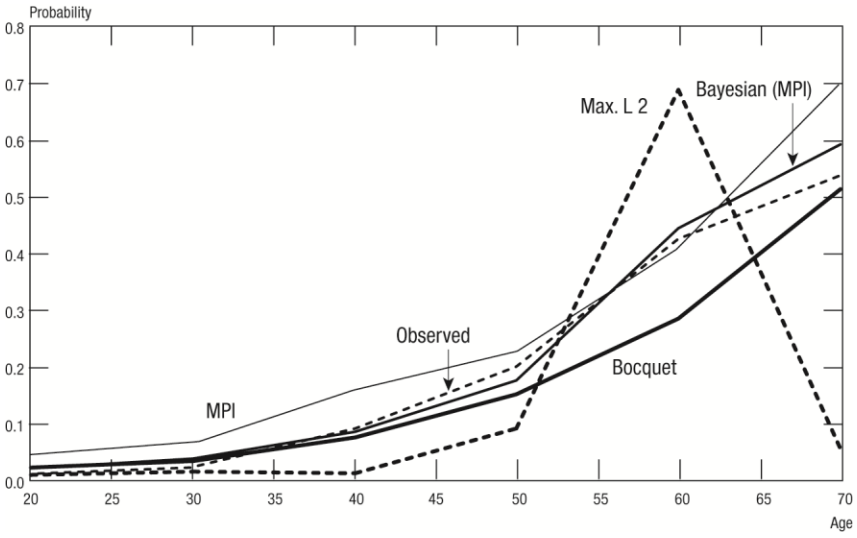
- for the Bayesian method: 0.003 and 0.040
- for the Bocquet-Appel and Bacro method: 0.014 and 0.078

demonstrating a significant advantage for our method.

It must be said, however, that if we had merely used the Bayesian analysis without taking account of the information provided by the particular situation of the nuns (our first analysis), we would have obtained distances 0.007 and 0.199. The method would have retained its advantage in crude deviation but lost it in relative deviation because of too large an "error" in the low probabilities.

We look now at the calculation of probabilities of death for the Maubuisson nun population. Figure 13.10 shows these probabilities calculated under various hypotheses.

Figure13.10. Probabilities of death for the Maubuisson nuns as estimated by the Bayesian method with prior from the modified pre-industrial standard (MPI), the method proposed by Bocquet-Appel and Bacro (Bocquet) and the Maximum Likelihood 2 method (Max. L. 2), compared with the modified pre-industrial standard (MPI) and the values actually observed (Observed) for these nuns.



This figure compares the ten-year probabilities of death estimated by the Bayesian method described here with as prior the modified pre-industrial standard and the method proposed by Bocquet-Appel and Bacro (2008) with the 756 vectors proposed by these authors (combination of Gompertz-Makeham distributions and extreme values - see previous chapter). For the purposes of comparison, the graph also shows the observed probabilities of death distribution for the nuns as estimated for the period 1640-1889. The IALK method (see previous chapter) gave results outside the limits [0, 1] for the proportion of deaths by age

group. This method can, however, be used with the addition of a positivity constraint. Here, we used the Maximum Likelihood 2 method (see Box 13.2).

It can be seen straightaway that this last method (Max. L. 2) provides rather unlikely probabilities of death close to zero except for the 50-60 and 60-70 age groups. Even if they are all positive, they are hardly acceptable. The Bocquet-Appel and Bacro method only gives probabilities of death close to the observed values for the first two age groups. They differ widely for the later groups, systematically underestimating the probabilities of dying. Conversely, the Bayesian method we propose gives quite accurate estimated probabilities for all ages, slightly overestimating for the first two age groups and underestimating for the last age group.

5. Conclusion

The previous chapter presented a detailed critical examination of the main approaches used by paleodemographers to estimate the age structure of a population for which they only have biological indicators measured from skeletons. Paleodemographers often call these methods Bayesian because they use Bayes' theorem and introduce a priori considerations into their method of estimation, but the paradigm upon which they are based is frequentist in nature. This chapter has rather proposed a strictly Bayesian approach in the sense habitually used in statistics, as specified in the introduction, in order to solve this major and recurrent problem for palaeographers. This conclusion provides an overview of the main advantages of this approach compared with the previous ones.

First, the previous approaches considered that the data taken from the observed groups (frequencies of data from reference population, frequencies of data from observed population) were entirely or partially fixed quantities. The probability vectors method and the IALK method took all these parameters as fixed in order to estimate the age structure of the observed population; the method proposed by Bocquet-Appel and Bacro (2008) considered the frequencies of data by stage taken from the observed population as fixed when establishing recommended confidence intervals. Since the numbers of skeletons, especially for the observed population, are often small, these hypotheses do not hold. The IALK method thus yields estimates that are incorrect or totally unrealistic (age-groups with zero probability) for the age structure of the observed population. The confidence intervals provided by Bocquet-Appel's Iterage software, which we ran with the Lisbon reference data for various site data frequencies, appear to be erroneous and much too small, and therefore over-optimistic: in some cases the interval is zero, some do not contain the estimated parameter value or that value is at one extremity of the interval. We shall see below a further possible reason for these inadequacies in the algorithm used.

Our approach considers all observed frequencies (both site and reference data) to be random; the same is true for the model in Appendix B , but we have seen that this in itself is not necessarily an advance. We continue by considering the unknown parameters (reference conditional probabilities and age probabilities) to be random under the Bayesian paradigm. Estimates are thus obtained in the form of posterior distributions of probabilities of various age classes, from which can be deduced point estimates (such as posterior means) and "Bayesian confidence intervals", more commonly known as "*credibility intervals*", whereas methods that do not allow for the randomness of the data cannot provide confidence intervals since these are based, by definition, on the uncertainty caused by the randomness.

Second, earlier approaches, except for that of Bocquet-Appel and Bacro, do not take into account the specific nature of the paleodemographic question, although the demographic

knowledge accumulated over many years and information about the living conditions of the populations concerned can provide information about their mortality that is potentially usable. Establishing networks of model life tables has made it possible to hypothesise a standard pre-industrial mortality that can be used to select a prior age distribution for the observed population that is more satisfactory than the uniform distribution. Similarly, in work on the Maubuisson nuns, say, one can use the fact that these women, because of their monastic lives, were not exposed to the same mortality risk as the general French population: in particular they avoided the risk of death in childbirth. Bocquet-Appel and Bacro's approach (2008) also uses prior information taken from paleodemographic research, but presented in a different manner. As we have said, the Iterage program introduces a restriction of the parametric space to make up for the small number of data. In some respects, their method is a regression, but instead of considering that the vector of the parameters to be estimated is in a space with as many dimensions as the number of age groups considered, with the sole restriction that its components sum to unity, it is confined within the convex envelope of vectors defined ex-ante by a mix of Gompertz-Makeham distributions and extreme values, for which the four parameters vary within restricted intervals. Although this mode of calculation is justifiable as it reduces the variability of the estimators, it may in return introduce a considerable bias. If the age distribution sought falls outside this restricted space, the estimate obtained may be at a considerable distance. This may result in some confidence intervals that never contain the true values, as pointed out above: this occurs if the intervals are confined within the same limits as the point estimates, as is the case with those provided by the Iterage algorithm.

The Bayesian approach differs in a third way from the IALK and frequentist approaches in general (except for the method proposed by Bocquet-Appel and Bacro, 2008). Whereas many authors have stressed that when the number of age groups is greater than the number of stages considered, no valid estimate of the age structure of the observed population is possible with a frequentist method, use of a Bayesian method removes this obstacle. We have seen that such an estimate is always possible in Bayesian terms; using examples we have even shown that the quality of the estimate may be improved with a finer division into age classes (see Séguy, Caussinus, Courgeau and Buchet [2012]).

A fourth feature sets the Bayesian approach apart. In the frequentist approach, it is usually considered sufficient to evaluate the mean and variance of the estimator of a parameter because, when the number of observations is large, its probability distribution often tends towards a normal distribution characterised by these two values. For the estimates in which we are interested (low probabilities and small samples), however, the distribution is generally highly asymmetric, casting doubt on this approach. The same asymmetry can be found in the posterior distributions of the Bayesian approach, but these distributions are easy to determine, making it possible to calculated reliable credibility intervals.

A fifth difference, this time between the Bayesian approach and the Iterage algorithm, is its simplicity of operation. Selecting a prior distribution is much easier and more flexible that the construction of a mortality model whose parameters are supposed to represent the most varied conditions of mortality, by attrition as well as disaster. Furthermore, once the paleodemographic sample differs from the model conditions, both in number of age classes and in selected intervals, the model has to be reset and all the prior probability vectors recalculated, whereas, to our knowledge, Bocquet-Appel and Bacro do not provide anywhere their full parametric formulation. It is clear that some situations will always fall outside such a mortality model, whereas the Bayesian method can be used to address all possible cases, albeit with varied efficacy.

In parallel to the "theoretical" considerations we have presented here, it is important to recall the results of our empirical studies. The simulations we have run can be used to measure the

quality of results obtained under the various methods proposed. Calculation of their mean squared errors provides a comparison between the estimates produced and the true values of the parameters, which in this case are known. The method we propose outperforms the other methods in almost all cases, often with substantial gains in accuracy. Only some of the results obtained with the Iterage algorithm manage to equal its performance, but in other cases Iterage introduces noticeable biases because it requires a restricted parametric space, whereas the age structure of the population under study falls outside that space.

We trust that this chapter has clearly demonstrated all the advantages of using a fully Bayesian estimation of the age structure of historic populations for which there is no record of age at death and where the records are replaced by the measurement of biological indicators. We hope that many paleodemographers will use it, providing further insight into its application and encouraging any improvements that may be necessary, so that their experience will complement our own.