Chapter 12
Critique of current methods

Daniel Courgeau[1]

To estimate the age structure of a population of skeletons, paleodemographers usually only have the structure by stage of biological indicators. As we shall see, these indicators alone are generally not sufficient for this estimation. So another source of information is needed to link the indicators to individuals' actual ages. It may be a *reference population* for which researchers have both types of measurement. The combination of the two ought to provide an age structure for the *observed population*. However, there are various possible solutions to the problem, and these have been extensively discussed among paleodemographers. In this chapter we examine these solutions and that discussion.

First, the names given to the methods used in paleodemography are often incorrect and contradictory. For example, what Masset, Bocquet-Appel, Jackes and others call the *vector method*, IPFP and IBFP[2] are referred to by Konigsberg, Frankenberg and others by such terms as the ALK and IALK[3] methods. Here they will be designated not by initials but names that clearly indicate their underlying hypotheses, because these methods have long been used in other disciplines and have a specific meaning that must be clearly defined. Obviously, the links between the more general name and the specific methods used by paleodemographers will be spelt out.

Second, we present in detail the aim of each method, independent of the discipline that uses it, in order to show its purpose. Each one will be applied to precisely identical examples so as to demonstrate its usefulness for paleodemography. We shall attempt to see which is the best suited to answering the questions the paleodemographer asks.

We shall start from the established hypotheses underpinning these estimates (biological uniformity), which are crucial for solving the problem posed. We shall, in particular, assume that the two sources mentioned above provide sufficiently reliable and fully usable information to estimate the age structure of buried populations. These hypotheses have been fully discussed throughout the book (see also Kemkes-Grottenthaler, 2002; Usher, 2002).

We set aside earlier estimates of less value to modern research: an excellent critical presentation of them can be found in Masset (1973). We shall only present the main two approaches currently used by most paleodemographers.

12.1 Tables of minimum distance between each cell

The principle of this method is to reconstitute the cells of a matrix knowing only its marginal values (the row totals and column totals) and then use an earlier reference matrix to improve the estimate. The criterion used is the greatest closeness between each cell in the reconstituted matrix and those of the earlier matrix. In paleodemography this means starting from the

[1] The author wishes to express his warmest thanks to Henri Caussinus for his highly relevant comments on the draft of this chapter, leading to a number of improvements. Thanks too go to Isabelle Séguy and Luc Buchet for their many comments on the same draft. The author is, of course, entirely responsible for the content.

[2] Iterative Proportional Fitting Procedure, Iterative Bayesian Proportional Fitting Procedure.

[3] Age Length Key, Iterated Age Length Key.

observed biological stages of a population under study and a *reference population* for whose members both age and biological stage are known and reconstituting the breakdown by age and stage of the population under study. This makes it possible to deduce that population's age structure.

12.1.1 Historical background

This approach was first used by Kruithof (1934), who was working on telephone networks. He wanted to start from a full telephone flowchart taken from a *reference population* and estimate a new matrix of telephone flows for an *observed population* where he knew only the marginal values (row and column totals).

This work was taken further in the 1940s (Deming and Stephan, 1940; Stephan[4], 1942) to estimate the cells of a contingency table subjected to a number of constraints on one or more of its marginal values, where all the cells of an initial table are known and are to be approximated as closely as possible. This method was used for census data, which were often incompletely tabulated.

It was later developed by Leontief (1941), Stone, Bates and Bacharach[5] (1963) in economics, Friedlander (1961), Thionet (1963, 1964), Caussinus (1965), Fienberg (1968, 1970) and Bishop *et al.* (1975) in statistics, Tugault (1970), Willekens (1977) and Willekens *et al.* (1981) in demography, etc., to be applied to increasingly complex cases, especially those where there was not even an initial matrix taken from a *reference population*. A detailed presentation is given in the chapter on estimates from incomplete data in Courgeau (1988), because these methods are widely used to estimate matrices of migration flows between zones, where information is incomplete (only the two marginal values known; earlier matrix known with, for example, only one or two marginal values for the period under study).

12.1.2. Table subject to constraints

This is a brief presentation of the simplest case where only the marginal values of the table are available and these are the constraints. The aim is to estimate all the cells without having a *reference population*, and then to do so with one.

Assume a two-entry matrix where only the marginal values are known (Table 12.1).

Table 12.1. Example of a simple matrix

		Age		Stage totals
		Group 1	Group 2	
Stage	Stage 1	m_{11}	m_{12}	3
	Stage 2	m_{21}	m_{22}	2
Age totals		4	1	5

It can easily be seen that if only the age and stage totals are known it is not possible to choose each intersecting age-stage cell arbitrarily. Here only two matrices are possible:

---

[4] In the second article, Stephan recognised that the results of the earlier one did not coincide from those of the least squares method, as Deming and Stephan (1940) had wrongly stated, but they did provide an approximate solution.

[5] These authors called it the RAS method, from the symbols they used in the input-output matrices. This name was common in the 1960s.

$$M_1 = \begin{pmatrix} 3 & 0 \\ 1 & 1 \end{pmatrix} \text{ and } M_2 = \begin{pmatrix} 2 & 1 \\ 2 & 0 \end{pmatrix}.$$

To choose between these two matrices, a further hypothesis must be made. Let us assume that each matrix corresponds to a macrostate generated by microstates induced by the deceased individuals themselves. Since we have five individuals i_1, i_2, i_3, i_4, i_5, it can be seen that the three individuals comprising m_{11} in M_1 can be chosen in a large number of ways:

$$(i_1, i_2, i_3) \text{ or } (i_1, i_2, i_4) \text{ or } (i_3, i_4, i_5) \text{ etc.}$$

If we then assume that each of these microstates is equally probable, the number of possible choices of three individuals for m_{11} is

$$\frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = \frac{120}{10} = 10$$

where 5! = 1 x 2 x 3 x 4 x 5 (factorial).

There then remain only two individuals, for whom the number of choices is simply 2. The last individual is necessarily counted in the last non-zero cell. There are therefore $\frac{5!}{3!2!}$ x 2 = 20 possibilities for five individuals to form matrix M_1. This is what is called the entropy of the macrostate.

Similarly, we show that there are

$$\frac{5!}{2!3!} \times \frac{3!}{2!1!} = \frac{5!}{2!2!1!} = \frac{120}{4} = 30$$

possibilities of forming matrix M_2 from five individuals. This is therefore the matrix we shall take as the most likely estimate for the matrix for which we know only the marginal values. In other words, we optimise the entropy.

It may be supposed that this method, where the two sets of marginal values are known, is also applicable in cases where, as in paleodemography, we only have the row marginal values (stage totals). In such cases there is more than one possible solution. For the example used here we find the same maximum number of possibilities for two different matrices:

$$M_1' = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \text{ and } M_2' = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$$

this number being $\frac{5!}{2!3!} \times \frac{3!}{2!1!} \times \frac{2!}{1!} = 60$ for the first matrix and $\frac{5!}{1!4!} \times \frac{4!}{2!2!} \times \frac{2!}{1!} = 60$ for the second. And yet the two matrices are quite different, since their column marginal values (age totals) are (3 2) and (2 3) respectively.

This shows clearly that, as we stated above, this method cannot provide an estimate of the age structure when all we have is the stage distribution of the *observed population*, except, of course, for the obvious case where it is known that a single age group corresponds to a single stage. In other cases, it is generally necessary to have a *reference population* whose age and stage structure is known. This brings us to the second case, in which we seek to come as close as possible to the reference.

12.1.3. Table subject to constraints and coming as close as possible to an initial table

Where the two sets of marginal values of an observed table are known, along with a reference table, the problem has been addressed in detail by many of the authors cited above, and leads to an iterative method, which has given it the universally accepted name of IPFP method. Where only one set of marginal values for the estimated table is known, the problem has received less attention, and we examine it here with simple examples.

12.1.3.1. Method used when only one set of marginal values is known

a. Example 1

This example starts from a matrix taken from a reference population giving the joint distribution of two characteristics:

$$M^0 = \begin{pmatrix} 40 & 10 \\ 20 & 30 \end{pmatrix}$$

If the observed population has still the same age totals as before, $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$, a new estimation of age structure may be obtained that includes this extra information. The principle is once again to estimate a matrix M whose terms come closest to those of matrix M^0 and which provides the age totals of the observed population.

The closeness between two matrices can, however, be measured in different ways. The simplest, on the face of it, is a Euclidean distance, the sum of the squares of the differences between each corresponding term in the two matrices divided by two. But this distance will depend largely on terms with a high value, where the differences may be greater than for terms of a low value. So that the difference may be best judged in relative terms. In this case, one can use a chi-square distance that will weight the squares of each difference by the numbers observed in the reference population. This distance[6], proposed by Deming and Stephan (1940) and Friedlander (1961), can be expressed as

$$d\left(M, M^0\right) = \frac{1}{2}\sum_{ij} \frac{\left(m_{ij} - m_{ij}^0\right)^2}{m_{ij}^0}$$

We shall seek to minimise it, subject to the constraints

$$m_{11} + m_{12} = 3 \text{ and } m_{21} + m_{22} = 2.$$

Because of these constraints, we shall use the Lagrange multiplier method. This calculates the partial derivatives of the distance, subject to constraints, with respect to each variable m_{ij}, which must be zero to obtain the minimum distance. This may be expressed, for example, for m_{11} and m_{12}:

$$\frac{m_{11} - m_{11}^0}{m_{11}^0} = \lambda \text{ or } m_{11} = m_{11}^0(\lambda + 1)$$

$$\frac{m_{12} - m_{12}^0}{m_{12}^0} = \lambda \text{ or } m_{12} = m_{12}^0(\lambda + 1)$$

---

[6] Minimising a chi-square distance involves additive adjustments whereas optimising an entropy involves multiplicative adjustments. The corresponding algorithms are very similar in spirit and results, but are, strictly speaking, different.

where $-\lambda$ is the Lagrange multiplier for the sum of the differentials $dm_{11} + dm_{12}$. Summing the two equations, the value of $\lambda+1$ may be estimated to be $\frac{m_{1.}}{m^0_{1.}} = \frac{3}{50}$. Applying this to the other two cells, we obtain this time a single solution:

$$m_1^{"} = \begin{pmatrix} \frac{12}{5} & \frac{3}{5} \\ \frac{4}{5} & \frac{6}{5} \end{pmatrix}.$$

This does give us the stage totals $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$, but the age totals are no longer integers (3.2 1.8). This is therefore a theoretical solution not verified in practice, since the numbers of individuals must be integers, but it may be considered as the age structure of a larger population, namely (0.64 0.36). The matrix obtained is the closest to the initial matrix in terms of chi-square distance. The use of other distances[7] would clearly lead to different estimates, which are generally not far from this one.

It can also be seen that this method requires only one iteration and that consequently the name Iterative Proportional Fitting Procedure is no longer really appropriate. We propose calling it simply Proportional Fitting Procedure (PFP).

b. Example 2

This example uses the paleodemographic data from Bocquet-Appel (2005, p. 297). It provides first a matrix of initial data where both individuals' age and the stage in which their femur is classified are known (Table 60).

We wish to estimate from this matrix the age structure of a new population (Loisy-en-Brie, Late Neolithic) for which we know only the femur stage distribution (Table 61). This is a clear case for a PFP procedure, with a reference matrix, Table 60, and only one set of marginal values for the observed population. Although the aim here is not to estimate the cells of the matrix, which are of no interest to the paleodemographer, but rather to estimate one set of marginal values from another, the method will still seek to find the closest matrix to Table 60, term by term, that provides stage totals equal to the numbers in Table 61.

The solution to this problem corresponds exactly to that presented above, minimising the chi-square distance when one set of marginal values for the table to be estimated is known and another matrix is available, estimated from a larger number of individuals (Deming and Stephan, 1940). In this case, we assume that the Loisy-en-Brie population is distributed among the osteological stages like the reference population: we work on row probabilities (the probabilities in each row sum to unity). This gives Table 62.

---

[7] Such as $d\left(M, M^0\right) = \sum_{ij} \frac{\left(m_{ij} - m^0_{ij}\right)^2}{s_{ij}}$, where $\frac{1}{s_{ij}}$ is the weight attached to m^0_{ij} (see Stephan, 1942), which may be measured, for example, by the variance of m^0_{ij}; $d\left(M, M^0\right) = \sum_{i,j} m_{ij} \ln \frac{m_{ij}}{m^0_{ij}}$ which is a measurement known as Kullback–Leibler divergence (also information divergence, information gain, relative entropy, or KLIC).

Applying these probabilities to each stage observed at Loisy-en-Brie, we obtain the following age at death structure:

$_{12}P_{23} = 0.244$, $_{12}P_{35} = 0.225$, $_{12}P_{47} = 0.222$, $_{12}P_{59} = 0.142$, $_{12}P_{71} = 0.132$, $_{12}P_{83} = 0.035$[8]

Once more it can be seen that where we have only one set of marginal values for the matrix to be estimated, a simple calculation with no iteration is sufficient, whereas with the standard IPFP method, where both sets of marginal values for the matrix to be estimated are known, more than one iteration is required.

Table 12.2 Population classified by age and femur stage (reference population)

Stages\Ages	23-34	35-46	47-58	59-70	71-82	83-94	Total
I	8	1	0	0	0	0	9
II	22	10	3	0	0	0	35
III	47	35	26	6	5	0	119
IV	13	29	35	30	25	5	137
V	1	4	10	10	9	4	38
VI	0	0	1	0	5	3	9
Total	91	79	75	46	44	12	347

Table 12.3. Numbers of femurs observed at Loisy-en-Brie, classified by stage*

Stages	Numbers
I	2
II	8
III	31,5
IV	40,5
V	12
VI	2
Total	96

*The numbers given by Bocquet-Appel for Stages III and IV are fractional, probably to allow for some approximate determinations.

Table 12.4 Transformation of Table 12.2 so that the probabilities in each row sum to unity.

[8] This notation may appear unusual but corresponds to the results previously presented in tabular form as below, or as a histogram.

Age group	23-34	35-46	47-58	59-70	71-82	83-94	Row total
Proportion	0.244	0.225	0.222	0.142	0.132	0.035	1

Stage\ages	23-34	35-46	47-58	59-70	71-82	83-94	Row sum
I	0,889	0,111	0	0	0	0	1
II	0,628	0,286	0,086	0	0	0	1
III	0,395	0,294	0,219	0,050	0,042	0	1
IV	0,095	0,212	0,255	0,219	0,182	0,037	1
V	0,027	0,105	0,263	0,263	0,237	0,105	1
VI	0	0	0,111	0	0,556	0,333	1

c. General case

Let n_{ij} be the number of individuals at stage i (out of a total of l stages) in age group j (out of a total of c groups), and $n_{i.}$ the various stage totals in the reference population. For the observed population, let π_i be its stage structure, the only known element. It can be shown that its age structure $\hat{p}_j$ is expressed by the formula

$$\hat{p}_j = \sum_{i=1}^{l} \pi_i \frac{n_{ij}}{n_{i.}} \qquad [12.1]$$

This corresponds to the matrix whose chi-square distance from the reference matrix is the least possible, subject to the numbers observed for each stage.

The results obtained with this method are identical to those from traditional age estimation in paleodemography, usually known as the probability vector method. This method goes back to Masset (1971) and was officially proposed by that author in 1973. Since then it has been used by many authors (Simon, 1982; Blondiaux, 1988; Pilet *et al.*, 1990; Danion *et al.*, 1994, Buchet, 1998). It is clearly described in Masset (1973, 1982, 1995) as the sum of the l vectors corresponding to the various probabilities for each individual in the *observed population* located at stage i of belonging to each age group j, based on probabilities calculated from the *reference population*. This sum must naturally be divided by the total number of individuals observed to obtain the age structure given by the formula above. So it is indeed the same principle that underlies both this method and the PFP method.

Now we shall show that this estimator $\hat{p}_j$ is also the same as that proposed by another approach.

12.1.3.2 The ALK method

The ALK (Age Length Key) method was initially proposed by Friðriksson (1934) for determining the age of all the fish caught of a given species from a random sample taken from the same catch. Although it is easy to classify the total population by length group, measuring

the age of each fish by closely observing their otoliths[9] is too expensive. So a small random number of fish are taken from each length group, the reference population, and their ages are accurately measured. The problem then is to estimate the age of fish taken from the total population on the basis of their length alone. The basic assumption of this method is that the fish in each length class in the reference population are a random sample of the observed population (Kimura and Chikuni, 1987). If so, then it is possible to estimate the age distribution of the entire observed population. This method is still used (Holden and Raitt, 1975; Farley and Basson, 2005) to estimate the age of fish caught. It has also been used in other fields, and in particular paleodemography.

It was the first method proposed by Konigsberg and Frankenberg (1992) under the same name, Age Length Key, to estimate the age distribution of past populations from a cross-distribution. But in this case we note that the basic assumption is quite likely not to hold, because we are working with two populations whose age structures have no particular reason to be the same.

More specifically, estimation by the ALK method involves first calculating a matrix of age group frequencies for each of the stages in the reference population. The numbers in each stage in the observed population are then attributed to the various age groups according to the frequencies calculated in the reference matrix for each of the observed stages. It only remains to sum the values in this matrix for each age group and divide each of these numbers by the total observed population to obtain the age distribution of the observed population. It can be seen that this amounts to calculating for the age group

$$\hat{p}_j = \sum_{i=1}^{l} \pi_i \frac{n_{ij}}{n_i}$$

This is the same estimator [12.1] as for the PFP method with chi-square distance, although it is based on apparently different principles. Naturally, if another distance is used, the two estimators will differ but generally only slightly.

12.1.4 Critique

Note that the distribution calculated in this manner necessarily depends on the age distribution of the reference sample and it is "flattened under the influence of the reference sample", as Masset (1995) puts it. This is a direct result of the assumption that each cell of the estimated matrix must be as close as possible to each cell of the reference matrix. Naturally, the greater the correlation between age and stage, the more satisfactory the estimate. Unfortunately, these correlations are rather slight in paleodemography, usually around 0.5 (Bocquet-Appel and Masset, 1982; Table 1, Chapter II.2 above), resulting in a major impact of the reference population on the age structure of the observed population.

Similarly, the assumption of the ALK method that the reference population is drawn from the observed population no longer holds even for two populations of the same fish species taken from different catches. This problem has been raised by a number of researchers (Kimura,

---

[9] Otoliths, literally "ear stones", are crystal structures found in the internal ears of fish and other vertebrates that are sensitive to gravity and linear acceleration. Their growth rings are used to estimate the fish's age.

1977) and is a crucial one in paleodemography, where the two populations are necessarily different, as mentioned above.

Note too that the two methods are used to estimate the theoretical observed matrix that is closest, term for term, to the initial matrix, which explains the dependency between the two matrices and ignores the invariance hypothesis (Müller *et al.*, 2002), also known as biological uniformity hypothesis (see Chapter I.3) under which for any human remains of a given age at death, the likelihood that it will be classified in a given stage only depends on that age, whatever population the bones or teeth have been taken from. Similarly, for fish, when the age structures of the reference population and the observed population are not necessarily the same, the more general hypothesis is made that the length structure for each age must be identical for the two populations. This hypothesis is in this case similar to the previous one, introducing a dissymmetry into the tables under consideration. It therefore becomes necessary to look for a more satisfactory method that takes full account of this hypothesis.

12.2 Tables of minimum distance between each column

Here it is not each cell in the reference matrix which must come as close as possible to the corresponding cell in the matrix of observed stages but each column with respect to its marginal value. In paleodemography, this means starting from the distribution by stage within each age group in the *reference population*. Then the weightings are found that, after multiplication by the various previously estimated distributions, give the numbers per stage in the *observed population*. The weightings will then correspond to the numbers per age in the *observed population*. In this case, the invariance hypothesis is perfectly verified. This problem, which differs from the previous one, must be solved by different methods.

12.2.1. Historical background

Here too the original task was to determine the age of an *observed fish population* for which only the length distribution is known. In this case, the *reference population* does not come from the same observed population but only a population of the same species, for which both length and age are known, as measured once again from the otoliths. Hasselblad (1966) provided an iterative method for an estimate of this type, followed by Orchard and Woodbury (1972), then Chikuni (1975). It was statistically developed by Kimura and Chikuni (1987), who proposed to call the method IALK, showing that it involved iterations. Unlike the ALK method, it only assumes that the length distributions for each age in the *reference population* are applicable to the *observed population*, which does not belong to the same total population (Kimura and Chikuni, 1987) and may therefore have a quite different age structure. These authors use the same algorithms and most of them use the term "mixture" for the method. It is in fact a special case of the more general likelihood maximisation (or expectation maximisation, EM) algorithm proposed by Dempster *et al.* (1977).

In his unpublished 1977 thesis, Bocquet-Appel proposed basing the estimate of the age structure of a population for which only the stage structure of its osteological remains are known on an iterative method starting from a uniform age structure. Masset (1982), in another unpublished thesis, proposed this *method of successive approximations* to avoid the exaggeratedly flat result of the *probability vector method*. To that end, he and Bocquet-Appel wrote an iterative program, called *Approx*, which he appended to his thesis. We shall see below that, when starting from a uniform distribution, this program gives the same results as the IALK method. But Masset wrongly states that this method "only gives really satisfactory results if the subjects tested are the same ones who were the basis of the sample" (Masset, 1982, p. 225), whereas in fact it avoids the need for that assumption, as we saw above. The application example for this method provided by Masset on pages 275-276 of his thesis, using

a population with seven age groups, leads to results it is hard to accept. Although he starts from a *reference population* with seven age classes and seven stages, and the vector of the stages in an *observed population* of 60 individuals contains no zero term, he obtains an age structure for the observed population that is rather unlikely:

(34.10 1.72 0 24.18 0 0 0),

because it contains four zero proportions. Faced with these disappointing results, he falls back on the less sophisticated method, thought more reliable, of *probability vectors*. In 1996, as we shall see, he adopted this method again with Bocquet-Appel.

Meanwhile, Konigsberg and Frankenberg (1992), looking for a more satisfactory method than ALK and wishing not to start from a uniform distribution like Masset and Bocquet-Appel, realised that the IALK method avoided these biases. First, where the ALK method gives good results when the reference population and the population whose age structure is to be estimated come from the same population, the IALK method avoids this constraint. The estimation can be based not on a uniform age distribution but any sort. Konigsberg and Frankenberg apply it to paleodemography using the maximum likelihood method (see Box 9 for the principle).

Bocquet-Appel and Masset (1996) continued with their *approximation method*, which they now, wrongly, called IPFP, explicitly referring on p. 572 to Deming and Stephan (1940). In the first part of this chapter, we showed that Bocquet-Appel and Masset sought to minimise the chi-square distance between each cell in the reference population matrix and the unknown matrix of the observed population, for which only one marginal value is known. IPFP is thus a misnomer, because the aim here is to minimise the distances between each of the columns corresponding to the same age: as we shall see, it is more like the IALK method. In order to distinguish it, we shall keep its original name of *approximation method*. In the same article, the authors also indicated the difficulties involved in getting this method to converge on acceptable results, and consequently proposed using it only to calculate the average age at death of individuals in the population. This restriction removes much of the method's usefulness.

These two approaches, which we shall call, for simplicity's sake, American and French, were the subject of considerable controversy between 1992 and 2002. But ultimately, they turn out to be practically identical (Konigsberg and Frankenberg, 2002; Konigsberg and Herrmann, 2002). Whereas the French insisted on starting from a uniform age distribution for their approach to give the right result, the Americans realised that their approach always gives the same result, whatever the initial distribution. We return to this point below. However, the French realised that the IALK method could lead to solutions with a large number of zero age groups, as we showed above in the presentation of Masset's thesis. This explains the highly sceptical attitude of French paleodemographers towards this method.

Box 12.1. Estimates by maximum likelihood

Henri Caussinus and Daniel Courgeau

Principle

In general, the statistical model states that data x are the observations of a random variable X, whose probability distribution depends on an unknown parameter θ. The probability density of X is a function of the data x of parameter θ, called likelihood function, often denoted L, and its value for x and θ, therefore, $L(x, \theta)$.

In frequentist (non-Bayesian) statistics, the most-used method is probably the so-called *maximum likelihood method*, which consists of estimating parameter θ by the value that maximises $L(x, \theta)$. The estimate is often denoted $\hat{\theta}(x)$, showing that it depends on observations x, or $\hat{\theta}$ in short.

In practice:

- Parameter θ can be of any value, but is usually (and, in the cases studied here, invariably) a real number or a family of k real numbers, with possible restrictions; for example, variance is necessarily positive or a probability between zero and unity.
- The data can be of any value, but are usually a list (vector) of real numbers; this may be a sample of n values x_i, observations of n independent random values X_i from the same distribution. The likelihood here is the product of the individual densities $f(x_i, \theta)$ and it is shown that, on very general hypotheses, the method is efficient for high values of n: see Part 2; but it is of much wider application, and although its efficiency cannot be guaranteed, one may hope that it still provide reasonable estimates.
- Where the data are discrete, the density $L(x, \theta)$ is understood as the probability of observation x.
- It is clear that one obtains the same estimate $\hat{\theta}(x)$ if $L(x, \theta)$ is multiplied by a quantity that does not depend on θ (but may depend on x); this will simplify the expressions in many cases (in formal terms, only the reference is being changed with respect to which the density is taken, an operation that must obviously be neutral for the estimate). Similarly, in finding the maximum, it is equivalent to replacing $L(x, \theta)$ by an increasing function of this expression, for example, its natural logarithm $\ln[L(x, \theta)]$.
- It is usual practice to find the maximum of a function by setting the derivative to zero, in the case of one variable, or the partial derivatives in a multi-dimensional case. This practice, which is justified for large samples on fairly broad hypotheses, does, however, require some precautions. First, the likelihood function may not be concave and the equations that set the derivatives to zero may have more than one solution and a choice must be made; however, this case will not be encountered in the questions covered by this book. Second, where the set of parameter variations is bounded, the maximum likelihood may be achieved on the frontier without corresponding to a zero-value derivative; this case is of direct interest to us, as we shall see in the example in Part 3.

Properties

The most significant properties of the maximum likelihood method of estimation are asymptotic properties concerning samples of n independent observations, where n tends to infinity; in practice, where n is sufficiently high. In this case,

$$L(x,\theta) = \prod_i f(x_i,\theta) \text{ and } \ln\left[L(x,\theta)\right] = \sum_i \ln\left[f(x_i,\theta)\right]$$

Moving on to the random variables, we have

$$\ln\left[L(X,\theta)\right] = \sum_i \ln\left[f(X_i,\theta)\right]$$

So $\ln[L(x, \theta)]$ is a sum of independent random variables to which, on very broad hypotheses, we may apply the general results of the calculation of probabilities such as the law of large numbers and the central limit theorem (normal distribution. We can then show that for high values of n

- the maximum likelihood is "regular", i.e., is obtained by setting at zero the derivative or derivatives of likelihood with respect to θ
- the probability distribution of $\hat{\theta}$ is approximately normal, centred on θ (absence of asymptotic bias), with a variance close to

$$\left\{ E\left[-\frac{d^2 \ln\left[L(X,\theta)\right]}{d\theta^2} \right] \right\}^{-1} \qquad [12.2]$$

in the case of a uni-dimensional parameter (E stands for expected value); in the multi-dimensional case, the matrix of second-order partial derivatives replaces the second derivative above to produce the matrix of variances and covariances of $\hat{\theta}$. Replacing θ by its estimate in this expression, we obtain an estimate of the variance of $\hat{\theta}$. So it is possible to give approximate intervals of confidence for θ or carry out tests. (Space does not permit demonstrations here, but note that the maximum of the likelihood function is less stable when it is poorly marked, i.e., the likelihood function is "flatter" near this maximum, its second derivative is smaller, making it intuitively likely that this derivative will appear in the variance of the estimator.)

3. Example

We may illustrate the point with a textbook example along the lines of problems of estimating the age distribution where the stage distribution is known, with a reference distribution assumed to be error-free (Example 1 in this chapter with two age groups and two stages). The reference probabilities are

$$\begin{matrix} 0.6667 & 0.25 \\ 0.3333 & 0.75 \end{matrix}$$

There is a single unknown parameter: let us take θ to be the probability p of the first age class. The stage probabilities are $0.6667p + 0.25(1 - p)$ and $0.3333p + 0.75(1 - p)$.

For m independent observations we have therefore the likelihood

$$L(x,p) = \prod_i \left[0.6667p + 0.25(1-p)\right]^{x_i} \left[0.3333p + 0.75(1-p)\right]^{(1-x_i)}$$

or

$$L(x,p)=[0{,}6667p+0{,}25(1-p)]^{m_1}[0{,}3333p+0{,}75(1-p)]^{m_2}=[0{,}25+0{,}4167p]^{m_1}[0{,}75-0{,}4167p]^{m_2}.$$

(In the first expression above, x_i is the indicator that is 1 if the observation is of stage 1 and 0 if it is of stage 2, and this formulation shows that it is a sample of size m; in the second expression, m_1 is the total number of observations of stage 1 and m_2 the total number of observations of stage 2, using the notation adopted in this book. One could also start from the binomial distribution of m_1, which would show in the likelihood a factor $m!/(m_1!m_2!)$, which, as we have seen, would have no effect.)

Figure 12.1 shows the variations of ln[$L(x, p)$] for two site data, respectively ($m_1 = 3$, $m_2 = 2$) and ($m_1 = 4$, $m_2 = 1$), the two cases considered in this chapter. It can be seen that if ($m_1 = 3$, $m_2 = 2$), there is a "regular" maximum; however, for ($m_1 = 4$, $m_2 = 1$) the regular maximum is "virtual", obtained for a value of p that is outside the range of possible values, since the "true maximum" for p is 1 and does not correspond to a zero value of the derivative.

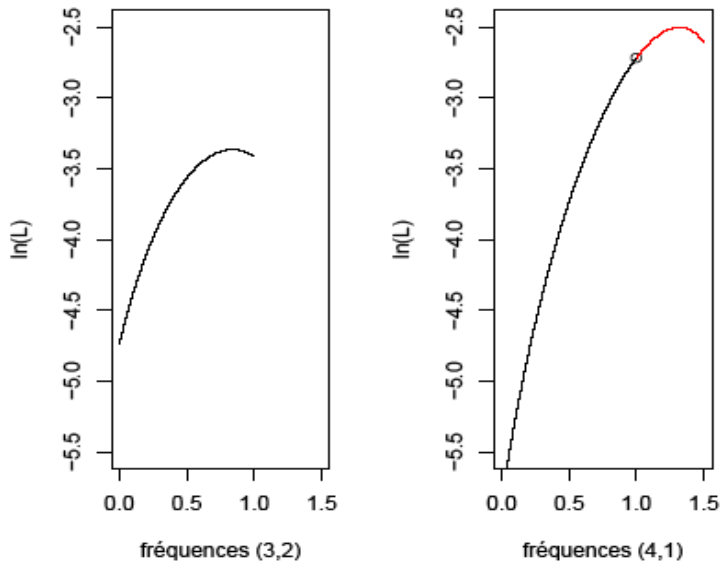


Figure 12.1 Log-likelihood for x = (3, 2), left, and ξ = (4, 1), right

Interpretation: although p on the x-axis can only vary between 0 and 1, the second curve has been extended to higher values (blue) to show the "virtual regular maximum".

For a sample of size m, we have

$$\ln[L(x,\ p)] = m_1 \ln(0.25 + 0.4167p) + m_2 \ln(0.75 - 0.4167p)$$

hence

$$\frac{d\ln[L(x,p)]}{dp}=\frac{0.4167m_1}{0.25+0.4167p}-\frac{0.4167m_2}{0.75-0.4167p}$$

which cancels to give $p = \frac{0.75m_1}{0.4167m} - \frac{0.25m_2}{0.4167m}$.

So we obtain for the two cases considered above $\hat{p} = 0.84$ and 1.32 respectively, values, it will be noted, that only depend on the relative frequencies m_1/m and m_2/m.

Let us also calculate expression [1]. Here we need to take the second derivative of the logarithm of likelihood expressed for the random values M_1 and M_2, whose observed values are m_1 and m_2. This gives

$$\frac{d^2 \ln\left[L(X,p)\right]}{dp^2} = -(0.4167)^2 \left[\frac{M_1}{(0.25 + 0.4167\,p)^2} + \frac{M_2}{(0.75 - 0.4167\,p)^2} \right].$$

Following the standard result from the binomial distribution, we have

$E(M_1) = m(0.25 + 0.4167p)$ and $E(M_2) = m(0.75 - 0.4167p)$, therefore:

$$E\left[-\frac{d^2 \ln\left[L(X,p)\right]}{dp^2} \right] = (0.4167)^2 m \left[(0.25 + 0.4167\,p)^{-1} + (0.75 - 0.4167\,p)^{-1} \right]$$

We obtain an approximation to the variance of the estimator of p by inverting this expression and replacing p by its estimate. Thus, for $m = 5$, $m_1 = 3$, $m_2 = 2$ (therefore $\hat{p} = 0.84$), as in the example above, we obtain 0.2765 as the approximation to the variance, hence a standard deviation of 0.526; this is high because of the small size of the sample, but it is important to remember that this value is extremely unreliable, since it is based on the assumption that the sample is a large one. If, however, the sample were larger, for example $m = 100$ with m1 = 60, $m_2 = 40$, the estimate of p would be the same but with an approximate variance one twentieth the size, namely 0.0138, with a standard deviation of 0.118, clearly more satisfactory and, not least, more reliable values.

We end by noting that if we had $m_1 = 80$, $m_2 = 20$ and a sample of 100, there would still be an "irregular" maximum likelihood for p equal to 1. However, with the values considered for the reference probabilities, to obtain $m_1 = 4$, $m_2 = 1$ for 5 observations is highly realistic, whereas $m_1 = 80$, $m_2 = 20$ for 100 observations is extremely unlikely because of the law of large numbers.

12.2.2. Maximum Likelihood Estimator

Take the general case with the same notation as above to formulate and solve the problem. Let us look first at the IALK method.

We apply the frequencies of the distribution of the biological indicator, conditioned by the age groups in the *reference population*, namely $f_{ij} = \frac{n_{ij}}{n_{.j}}$. For the *observed population*, we have stage frequencies π_i. Applying the maximum likelihood method shows that the age structure $\hat{p}_j$ may be obtained by successive iterations from any initial structure, which is often taken to be $\frac{1}{c}$, i.e., uniform:

$$\hat{p}_j^{n+1} = \sum_{i=1}^{l} \pi_i \frac{\hat{p}_j^n f_{ij}}{\sum_{j=1}^{c} \hat{p}_j^n f_{ij}} \qquad [2]$$

As many iterations as necessary are run for $\hat{p}_j^n$ to differ from $\hat{p}_j^{n+1}$ by as small a quantity as desired. An estimate is thus obtained of the age structure of the observed population. Konigsberg and Frankenberg (2002) recognise that the estimators they obtain in this way are estimators of maximum likelihood and not Bayesian estimators. Furthermore, this solution is only valid if the estimators are all positive: if some are zero, this may perhaps not be a maximum likelihood solution. It is also possible to estimate the variances of these estimators, as Cribari-Neto and Zarkos (1999) have done. We shall not develop these estimates here but shall use them below to estimate their standard deviation.

This method is easy to generalise to the case where there are more than one biological indicators or where these indicators are not discrete but continuous (Konigsberg and Frankenberg, 1992).

It can also be used to introduce a continuous, rather than discretised, age without changing the principle. This is well laid out in Hoppa and Vaupel's edited volume (2002a), following a workshop on the topic at the Max Planck Institute for Demographic Research in Rostock, attended by a large number of English-speaking anthropologists, but with no French specialists invited. Konigsberg and Herrmann (2002) clearly indicate the closeness of the results obtained by IALK and these more sophisticated methods: "Our current methods fit fairly comfortably within the approaches taken during the Rostock workshop"[10]

First, the age distribution of a given stage in the reference population – with age now treated as a continuous variable – is provided by various types of non-parametric or parametric regression models. However, the volume's main originality is the use of a parametric event-history model (Courgeau and Lelièvre, 1989) to model the probability density of the observed population's mortality. Provided the model does not include too many parameters (Gompertz two-parameter model, Gompertz-Makeham three-parameter model, Siler five-parameter model, etc.), we can estimate it using the maximum-likelihood method with the previously estimated age distribution of stages. Applying a notation similar to the previous one, we can summarise this formulation in the following form, where the age variable j is now continuous:

[10] The initial workshop was held in June 1999.

$$\pi_i = \int_j w_i(j) p(j,\theta) dj$$

and where $w_i(j)$ is the distribution of the stage i by age j, estimated in the *reference population* and $p(j, \theta)$ the age-specific probability density of the *observed population*, whose parameters θ we need to estimate using the l similar relations for each stage.

The problem is that these methods introduce a number of additional hypotheses, notably: a stationary or stable population, so that the even-history model can apply to a current population; and continuity in the age distribution of a given stage, yielding different estimates according to the methods used. In principle, therefore, there is no reason why these hypotheses – which we have no way of verifying – should be fully satisfactory. For instance, a past population that has experienced an epidemic cannot be considered stationary or stable. Similarly, to impose on that population a parametric event-history model – ultimately rather simple and verified on current populations – may fail to capture situations where these models were not verified. Not least, these methods still assume that the *reference population* is perfectly observed although sampling errors in paleodemography can be considerable. Not to allow for that fact, as with the IALK method, introduces a major risk of error into the estimate of the age structure of the *observed population*.

12.2.3. Approximation method

The *approximation method* proposed by French-speaking authors is also iterative, but presented in a more experimental than truly mathematical manner. Bocquet-Appel and Masset (1996) contains a detailed description of the procedure, where both p_j and f_{ij} depend on the iteration. We shall show that in fact it is not necessary for them to depend on the iteration, and in this case the method can, in some conditions, lead to the same result as IALK.

It starts from an initial value for the age breakdown which is assumed from the outset to be uniform, $\frac{m}{c}$, where the total number of the observed population m is distributed among the c age groups. The two important relationships for this algorithm are the following

$$\hat{p}_j^n = \sum_{i=1}^{l} \pi_i \frac{\hat{f}_{ij}^{n-1}}{\sum_{j=1}^{c} \hat{f}_{ij}^{n-1}} \text{ and } \hat{f}_{ij}^n = \hat{f}_{ij}^{n-1} \frac{\hat{p}_j^n}{\hat{p}_j^{n-1}}$$

Starting from the initial value $\hat{p}_j^0 = \frac{m}{c}$ and the initial value $f_{ij}^0 = f_{ij}$, we deduce from the first relationship

$$\hat{p}_j^1 = \sum_{i=1}^{l} \pi_i \frac{f_{ij}}{\sum_{j=1}^{c} f_{ij}} \qquad [12.4]$$

Note that in this first iteration the formulation [12.4] differs from the general formulation [12.3]. The second relationship then gives

$$\hat{f}_{ij}^1 = \frac{c}{m} f_{ij} \hat{p}_j^1$$

and finally, by the first relationship:

$$\hat{p}_j^2 = \sum_{i=1}^{l} \pi_i \frac{\hat{p}_j^1 f_{ij}}{\sum_{j=1}^{c} \hat{p}_j^1 f_{ij}}$$

Now we are back at formula [12.3]. It can also be seen that there is no point in considering $\hat{f}_{ij}^1$ or the number $\frac{m}{c}$, which cancels out from top and bottom. It only remains to see whether

$$\hat{f}_j^{n-1} = \frac{c}{m} f_{ij} \hat{p}_j^{n-1}$$

is verified, so

$$\hat{p}_j^n = \sum_{i=1}^{l} \pi_i \frac{\hat{p}_j^{n-1} f_{ij}}{\sum_{j=1}^{c} \hat{p}_j^{n-1} f_{ij}} \text{ and } \hat{f}_{ij}^n = \frac{c}{m} f_{ij} \hat{p}_j^n$$

which is easily shown by using the previous algorithm. As we have demonstrated that these relationships held for $n = 2$, they therefore hold for all values of n. Again we are back at the same formula [12.3] as for IALK, from the second iteration on. What Konigsberg and Frankenberg showed empirically has thus been proven mathematically in its most general form.

However, the main difference between the two methods is that the first can be used with any initial structure, as long as its age values sum to unity, whereas the second one requires starting from a uniform structure. This is simply due to the formulations and consequent different values for the first iteration, because, from the second iteration on, the formulations are identical. If we take a non-uniform initial distribution for the second method, the solutions found will no longer be maximum likelihood estimators.

In the first example, it can be seen that, starting from a uniform initial age distribution (2.5 2.5), one arrives at the solution (0.84 1.16) after some 70 iterations, as with the IALK method. However, as soon as the initial distribution is slightly different, the solution becomes widely different. For example, for an initial age group with numbers varying by 2.35 to 3.25, the estimate of the final population for that group will fall from 5 to 0. Below 2.35, the final population of the first age group will remain at 5, and above 3.25, it will be zero. It can be seen how sensitive the final estimate is to the initial distribution, which must be taken to be precisely uniform.

It is instructive to use the same *reference population* for a larger *observed population* comprising 50 skeletons, for example. As long as the observed population at the first stage is fewer than 13, the population in the first age group is zero. But, as soon as the population at that stage exceeds 13, the estimated population in the first age group increases regularly from 0 to 50, when the population at this stage reaches 34 individuals. It can be seen that the structure of the *reference population* can be used to estimate all possible combinations by age for the *observed population*. This contradicts the idea that some structures cannot be found with this method, as claimed by Bocquet-Appel (2005), as we shall see below. Naturally, this case needs to be extended to a larger number of age groups, but then the number of possible combinations becomes too high for the extension to be done properly. However, it can be clearly seen that in some cases we always find a population for which only one age group is represented. This coincides with Masset's result (1982), mentioned above, where certain age groups are estimated at zero.

12.2.4 Summary of the two methods

Now we examine the more general case of an observation of l stages in a population with c age groups. The reference population is given in Table 12.5, numbers n_{ij} by age j and stage i.

We seek to estimate the age structure of a new population for which only the numbers by stage m_i are known, given in Table 12.6.

Unlike the PFP estimate, where the probabilities to be estimated make no difference between rows and columns, this method considers the initial reference table in an asymmetric fashion. The invariance hypothesis involves assigning a clear significance to the probability of belonging to stage i if we are dealing with a given age group j, which we shall denote $p_{i|j}$, which is supposed to be applicable to any observed population. These probabilities then verify the relationship:

$$\sum_j p_{i|j} = 1 \text{ for all values of } j.$$

To estimate these probabilities from Table 63, we shall see below the various assumptions that may be made.

Table 12.5. Reference population matrix by stage and age group

	Age group j									Stage totals
Stage i	n_{11}	.	.	.	n_{1j}	.	.	.	n_{1c}	$n_{1.}$
	.									
	.									
	.									
	n_{i1}	.	.	.	n_{ij}	.	.	.	n_{ic}	$n_{i.}$
	.									
	.									
	.									
	n_{l1}	.	.	.	n_{lj}	.	.	.	n_{lc}	$n_{l.}$
Age totals	$n_{.1}$	.	.	.	$n_{.j}$	.	.	.	$n_{.c}$	$n_{..}$

Table 12.6. Size of the observed population by stage

Stage totals	m_1	.	.	m_i	.	.	m_l	m

For the observed population, we need to estimate the probability of belonging to one of the various stages p_i from the observed frequencies in Table 64. Using these probabilities and those estimated from Table 63, we can then estimate the age structure of this population $p_{.j}$, which we are seeking.

The initial matrix is used to calculate for each age group j a vector giving the distribution frequencies of biological indicator i:

$$f_{ij} = \frac{n_{ij}}{n_{.j}}$$

If we assume that the observed numbers are high enough, the conditional age probabilities for a given stage will equal these frequencies $p_{i|j} = f_{ij}$. This gives us c vectors, for which we take no account of any sampling errors in the reference population.

For the new observed population, we assume once again that the observed numbers are high enough. This gives us a vector estimating the probabilities of an individual from this population belonging either to the various stages π_i or for stage i:

$$\pi_i = \frac{m_i}{m}$$

From these estimates we can devise a method for estimating age structures.

12.2.4.1 Devising the model

We wish to know if there is a set of weightings $(p_1, p_2, \cdots, p_c)$ representing the required age structure such that they verify the relationships:

$$\sum_j p_j f_{1j} = \pi_1$$

.

$$\sum_j p_j f_{ij} = \pi_i \qquad [12.5]$$

.

.

$$\sum_j p_j f_{lj} = \pi_l$$

A final condition must be added:

$$\sum_j p_j = 1$$

which is a necessary one, because the weightings must sum to unity. So it is a mixture of distributions, *prior* data, we must estimate to find the distribution by stage of the observed population.

12.2.4.2 Solution for a square matrix

If the matrix is square ($l = c$), Cramer's rule applies, with generally one and only one solution. It is easy to see that the additivity condition for the p_j to be equal to one is necessarily satisfied.

This application of Cramer's rule can be expressed more simply in matrix notation.

If F is the square matrix

$$F = \begin{pmatrix} f_{11} & . & f_{1i} & . & f_{1c} \\ . & & & & . \\ f_{i1} & . & f_{ii} & . & f_{ic} \\ . & & & & . \\ f_{l1} & . & f_{li} & . & f_{lc} \end{pmatrix}$$

π the column vector

$$\pi = \begin{pmatrix} \pi_1 \\ . \\ . \\ \pi_l \end{pmatrix}$$

and p the column vector

$$p = \begin{pmatrix} p_1 \\ . \\ p_j \\ . \\ p_c \end{pmatrix}$$

The previous system of equations can be expressed compactly as

$$Fp = \pi$$

The estimator of p, $\hat{p}$, is then obtained by multiplying to the left the two members by the inverse of the F matrix, which is generally calculable:

$$\hat{p} = F^{-1}\pi$$

When the matrix is not square:

if $l < c$, the system is indeterminate and admits an infinity of solutions.

if $l > c$, the system usually has no solution, but it is possible to calculate the $\hat{p}_j$ by using statistical methods.

12.2.4.3 Least squares method

This involves seeking those values of p_j that minimise the following sum of squares:

$$S = \left(\sum_j p_j f_{1j} - \pi_1 \right)^2 + ... + \left(\sum_j p_j f_{ij} - \pi_i \right)^2 + ... + \left(\sum_j p_j f_{lj} - \pi_l \right)^2$$

with the constraint $\sum_j p_j = 1$. It can be seen, first, that where $l = c$, the solution $Fp = \pi$ given for a square matrix is always verified, because the solution is the same for both methods.

Where $l > c$, if we assume that we have values for p_j, then by introducing a variation ∂p_j, the differentials of the two previous equations give:

$$\tfrac{1}{2}\partial S = \left[f_{11}\left(\sum_j p_j f_{1j} - \pi_1\right) + \ldots f_{l1}\left(\sum_j p_j f_{lj} - \pi_l\right)\right]\partial p_1 + \ldots + \left[f_{1c}\left(\sum_j p_j f_{1j} - \pi_1\right) + \ldots f_{lc}\left(\sum_j p_j f_{lj} - \pi_l\right)\right]\partial p_c = 0$$

$$\text{and } \sum_j \partial p_j = 0$$

Now, multiplying the last equation by the arbitrary Lagrange multiplier λ and adding the two equations, we obtain:

$$\left(\frac{\partial S}{\partial p_1} + \lambda\right)\partial p_1 + \ldots + \left(\frac{\partial S}{\partial p_c} + \lambda\right)\partial p_c = 0$$

leading to a linear system for $(c + 1)$ equations with $(c + 1)$ unknowns $p_1, p_2, \ldots p_c, \lambda$:

$$\begin{cases} p_1\sum_i f_{i1}^2 + \ . \ . \ + p_j\sum_i f_{i1}f_{ij} + \ . \ . \ + p_c\sum_i f_{i1}f_{ic} + \lambda = \sum_i f_{i1}\pi_i \\ \quad . \qquad . \ . \qquad . \qquad . \ . \qquad . \qquad . \\ \quad . \qquad . \ . \qquad . \qquad . \ . \qquad . \qquad . \\ p_1\sum_i f_{ic}f_{i1} + \ . \ . \ + p_j\sum_i f_{ic}f_{ij} + \ . \ . \ + p_c\sum_i f_{ic}^2 + \lambda = \sum_i f_{ic}\pi_i \\ \quad p_1 + \qquad . \ . \qquad + p_i + \qquad . \ . \qquad + p_c \qquad\qquad = 1 \end{cases}$$

If the *invariance* condition and the assumption that the observed numbers are not subject to uncertainty are verified, we should find as solution to this system of equations by Cramer's rule a system of weightings all of whose values lie within [0, 1], corresponding to the structure of age at death. But, since the data are necessarily subject to uncertainty, because there are few of them, and since this estimate is obtained by least squares, it may be that some estimates fall outside [0, 1], even if the invariance hypothesis is verified. If so, the problem of estimating the structure of age at death will need to be solved allowing for these measurement errors, as we shall see below.

12.2.4.4 Maximum likelihood method

This method consists of considering that system [12.5] gives the probability π_i that an individual belongs to stage i. It is then possible to calculate the probability of observing a sample of m individuals of whom m_1 are at stage 1, m_2 at stage 2, etc., which will constitute its likelihood:

$$\frac{m}{\prod_{i=1}^{l} m_i!}\prod_{i=1}^{l}\left(\sum_{j=1}^{c} p_j f_{ij}\right)^{m_i}$$

with, as before, the constraint:

$$\sum_{i=1}^{l} m_i = m$$

Since the first fraction of likelihood is independent of the p_j, it suffices to maximise the logarithm of the second expression with the constraint. If this maximum satisfies the p_j positivity constraint, it will be found by setting at zero the likelihood gradient. Using the Lagrange multiplier, we then obtain the system of the following c equations to be solved:

$$\sum_{i=1}^{l} \frac{m_i f_{ij}}{\sum_{j=1}^{c} p_j f_{ij}} - \lambda = 0$$

It can easily be seen that by multiplying each equation by p_j and summing the l equations, we obtain $\lambda = m$. The result is the non-linear system of c equations with c unknowns:

$$\sum_{i=1}^{l} \frac{\pi_i f_{ij}}{\sum_{j=1}^{c} p_j f_{ij}} - \lambda = 1$$

It can be seen that where $l = c$, the solution $Fp = \pi$ is always verified and that this method gives the same result as the least squares method.

It can be demonstrated that this system can be solved by the following iterations (Hasselblad, 1966):

$$p_j^n = \sum_{i=1}^{l} \frac{\pi_i p_j^{n-1} f_{ij}}{\sum_{j=1}^{c} p_j^{n-1} f_{ij}}$$

which are the same as those in formula [12.3]. Starting from any positive values of p_j^0 that sum to unity, and setting a threshold beyond which the values of p_j^{n-1} and p_j^n are taken to be equivalent, we obtain the solution of the system by the maximum likelihood method.

It can be seen that these iterations are identical to those proposed by such paleodemographers as Konigsberg and Frankenberg (2002) and Bocquet-Appel and Masset (2005). Note, however, that since it starts from positive values, this algorithm can only give positive or zero values. If the point of zero gradient on the likelihood curve does not fall within the domain of possible values, the IALK algorithm gives an estimate on the frontier of this domain (at least one of the p_j is zero), for which it has not been proven that it is a maximum likelihood for $c > 2$ (although it is probably true in many cases). For that purpose, in the following chapter, the maximum likelihood will be found by a procedure other than the IALK algorithm.

The least squares and maximum likelihood methods lead to different solutions where $l > c$, but these are generally close together. It is instructive to compare them in the application examples.

12.2.4.5 Application examples

We shall now apply the two methods to four populations to see if they are always valid for paleodemography.

a. Theoretical example 1: 2 age groups, 2 stages

We use once more the earlier simplified example, which provides, as we have shown, the age structure (0.84 0.16). Since $l = c = 2$, it is easy to verify that the solutions given by the least squares and maximum likelihood methods are both identical to that given by direct calculation.

Note, however, that it is not possible to start from any marginal row value to find a column structure that fits with the two stage structures given for each age. It can easily be seen that the proportion for stage 1 of the observed population must fall between $\frac{1}{4}$ and $\frac{2}{3}$, in order to

obtain solutions within the limits [0, 1] of the probabilities. In this particular case, this is so, because $\frac{3}{5}$ does lie in this interval $[\frac{1}{4}, \frac{2}{3}]$.

However, if we had an observed population with stage values of $\begin{pmatrix} 4 \\ 1 \end{pmatrix}$, which is perfectly possible if we observe, still from the same larger overall population, a sample of only five individuals, then the age structure estimated by the analytical method will be (1.32 – 0.32). It can be seen that the two solutions always sum to unity, but that they do not correspond to probabilities. Interestingly, the IALK method leads to a different solution (1 0), which does correspond to the maximum likelihood, but a maximum on the frontier of possible values, where the derivative of likelihood is not zero (see box 12.1). In practice, we have here a case where none of the proposed solutions is appropriate.

b. Theoretical example 2: 2 age groups, 3 stages

Now we take the same number of age groups and one extra stage. Here we shall estimate the least squares and maximum likelihood solutions.

The reference population, established for this example, gives us the following matrix:

$$\begin{pmatrix} 40 & 10 \\ 20 & 30 \\ 4 & 40 \end{pmatrix}$$

and for the population whose age structure is to be estimated we have the following breakdown by stage:

$$\begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

From this we deduce the two stage structures of the reference population for each age group:

$$S_1 = \begin{pmatrix} 0.625 \\ 0.3125 \\ 0.0625 \end{pmatrix} \text{ and } S_2 = \begin{pmatrix} 0.125 \\ 0.375 \\ 0.5 \end{pmatrix}$$

and the stage structure of the population for which we wish to estimate the age structure:

$$\pi = \begin{pmatrix} 0.5 \\ 0.3333 \\ 0.1667 \end{pmatrix}$$

If we apply the least squares method, we obtain the following system of equations:

$$\begin{cases} 0{,}49219p_1 + 0{,}22656p_2 + \lambda = 0{,}42700 \\ 0{,}22656p_1 + 0{,}40625p_2 + \lambda = 0{,}27087 \\ p_1 + p_2 = 1 \end{cases}$$

This system is solvable and leads to the age structure of the observed population (0.7549 1.2451), with a parameter λ at the very low value of 0.0004.

If we now apply the maximum likelihood method, we obtain the following age structure (0.7576 0.2424), which is very close to the preceding one.

c. Example 3: Loisy-en-Brie population

Here we take the example given by Bocquet-Appel (2005) of the Loisy-en-Brie population estimated from a reference population for which both age of individuals and stage in which their femurs are classified are known (Table 12.2).

Here we have a case where $l = c = 6$. The analytical solution is therefore applied, without turning to the least squares or maximum likelihood methods. The matrix F used is given in Table 12.7, where the column probabilities now sum to unity.

Table 12.7. Transformation of Table 12.2 so column probabilities sum to unity

	23-34	35-46	47-58	59-70	71-82	83-94
I	0.088	0.013	0.000	0.000	0.000	0.000
II	0.242	0.126	0.040	0.000	0.000	0.000
III	0.516	0.443	0.347	0.131	0.114	0.000
IV	0.143	0.367	0.467	0.652	0.568	0.417
V	0.011	0.051	0.133	0.217	0.204	0.333
VI	0.000	0.000	0.013	0.000	0.114	0.250
Total	1.000	1.000	1.000	1.000	1.000	1.000

We calculate the inverse of the matrix:

$$\begin{pmatrix} 16{,}996 & -3{,}015 & 0{,}500 & -0{,}186 & 0{,}259 & -0{,}035 \\ -39{,}029 & 20{,}934 & -3{,}476 & 1{,}294 & -1{,}798 & 0{,}240 \\ 20{,}784 & -23{,}024 & 7{,}974 & -2{,}970 & 4{,}125 & -0{,}550 \\ -7{,}814 & 9{,}650 & -3{,}358 & 0{,}140 & 6{,}195 & -8{,}493 \\ 20{,}481 & -8{,}750 & -0{,}395 & 4{,}699 & -13{,}861 & 10{,}648 \\ -10{,}417 & 5{,}205 & -0{,}246 & -1{,}978 & 6{,}080 & -0{,}810 \end{pmatrix}$$

which, multiplied by the vector of the probabilities of belonging to each stage obtained from Table 12.2 finally gives the structure by age at death to be estimated:

$${}_{12}p_{23} = 0.219,\ {}_{12}p_{35} = 0.119,\ {}_{12}p_{47} = 0.380,\ {}_{12}p_{59} = 0.197,\ {}_{12}p_{71} = 0.040,\ {}_{12}p_{83} = 0.045.$$

Since $l = c = 6$, this solution is always identical to that obtained by the maximum likelihood or least squares methods.

However, these solutions differ sharply from that obtained by the PFP method:

$${}_{12}p_{23} = 0.244,\ {}_{12}p_{35} = 0.225,\ {}_{12}p_{47} = 0.222,\ {}_{12}p_{59} = 0.142,\ {}_{12}p_{71} = 0.132,\ {}_{12}p_{83} = 0.035$$

This last method provides an age distribution closer to the reference population than the previous one.

d. Example 4: the Maubuisson nuns (17th-18th centuries)

Here we use cranial suture closure as the age indicator for the female reference population established in Chapter 4, for which we have combined certain ages and stages (Table 12.8) so as not to take up too much space, but the result is the same when all values are considered.

Table 12.8. Distribution by stage (combined suture closure coefficients) and age group observed in the female reference population Preference (established from three Portuguese collections, see Chapter 4)

Stages\ages	20-29	30-39	40-49	50-59	60-69	70-79	80+	Total
0-4	85	40	45	26	11	7	5	219
5-7	6	13	12	10	2	4	5	52
8-12	6	11	15	14	6	11	5	68
13-18	5	2	11	13	13	11	11	66
19-23	3	6	6	11	7	12	8	53
24-30	3	5	6	12	19	13	12	70
31-40	1	6	2	14	12	8	23	66
Total	109	83	97	100	70	66	69	594

The observed population is distributed by stage[11] (Table 12.9).

Table 12.9. Distribution by stage observed in the archaeological population

Stages/numbers	Population
0-4	6
5-7	2
8-12	4
13-18	5
19-23	3
24-30	9
31-40	8
Total	37

Once again it can be seen that, since the matrix is square, the analytical method is applied. The corresponding matrix F is

[11] The stage division adopted here is slightly different from that used in Chapter III.1.2., since there needed to be at least as many stages and age groups (7).

$$\begin{pmatrix} 0{,}780 & 0{,}482 & 0{,}464 & 0{,}260 & 0{,}157 & 0{,}106 & 0{,}073 \\ 0{,}055 & 0{,}157 & 0{,}124 & 0{,}100 & 0{,}029 & 0{,}060 & 0{,}073 \\ 0{,}055 & 0{,}133 & 0{,}155 & 0{,}140 & 0{,}086 & 0{,}167 & 0{,}073 \\ 0{,}046 & 0{,}024 & 0{,}113 & 0{,}130 & 0{,}186 & 0{,}167 & 0{,}158 \\ 0{,}028 & 0{,}072 & 0{,}062 & 0{,}110 & 0{,}100 & 0{,}182 & 0{,}116 \\ 0{,}028 & 0{,}060 & 0{,}062 & 0{,}120 & 0{,}271 & 0{,}197 & 0{,}174 \\ 0{,}008 & 0{,}072 & 0{,}020 & 0{,}140 & 0{,}171 & 0{,}121 & 0{,}333 \end{pmatrix}$$

Omitting intermediate calculations, we arrive at the structure of age at death to be estimated:

$$_{10}p_{20} = 0.091,\ _{10}p_{30} = -0.505,\ _{10}p_{40} = -5.018,\ _{10}p_{50} = 11.560,$$

$$_{10}p_{60} = 1.709,\ _{10}p_{70} = -3.387,\ _{10}p_{80+} = -3.450$$

Although these figures sum to unity, they cannot now be used as a weighting system, since some are negative and others exceed unity. The numbers observed here are too small to give an age structure with all values lying within [0, 1]. The analytical method cannot provide an acceptable estimate here. Naturally, a positivity constraint could be introduced for the probabilities, but in that case some estimates would be zero, which would still not be satisfactory.

Whereas the least squares method always gives an identical solution, since $l = c = 7$, the maximum likelihood method, with its standard algorithm, leads to the following solution:

$_{10}p_{20} = 0,\ _{10}p_{30} = 0.110,\ _{10}p_{40} = 0,\ _{10}p_{50} = 0,\ _{10}p_{60} = 0.614,\ _{10}p_{70} = 0,\ _{10}p_{80+} = 0.276.$

As can be seen, this algorithm, which imposes positive probabilities, leads to a totally different estimate for the age structure, with four age groups with zero values. Once again, the solution is not acceptable.

However, in this case, where $l = c$, it is possible to estimate the maximum likelihood solution directly, because we obtain a square matrix that can be inversed. It can be verified that the age structure is the same as that obtained by the analytical solution or the least squares method.

12.2.5 Critique

First, some authors use this method where $l > c$. In some cases the algorithm appears to converge towards a solution, but the variance of the estimates increases excessively and removes any interest the method might have. For example Bocquet-Appel and Bacro (1997) apply this method to estimate seven age classes although they have only six stages. Although the iteration results seem to converge properly, Konigsberg and Frankenberg (2002) estimate that the standard deviations of these estimators are respectively

(576 3.156 9.496 14.355 25.786 22.084 3.737),

demonstrating that the results are incoherent. An even more extreme example is Jackes (2000) attempting to estimate 17 age classes with only 6 stages. Her results include a large number of zero-value age classes, clearly indicating that the model has not been identified. Furthermore, since this iteration procedure cannot provide negative probabilities, which should happen here, the positive results do not even verify the conditions in which the columns of the estimated matrix come closest to those of the reference matrix. Konigsberg and Frankenberg (2002) clearly state the conditions required to obtain a solution, as we have done above: the

number of stages must be equal to or exceed the number of age groups considered. This condition is not verified in all the preceding examples, which explains the incoherencies.

Whereas Konigsberg and Frankenberg (1992) consider that they can correctly estimate the age structure of the observed population by this method, Bocquet-Appel and Masset (1996) believe that an estimate of this sort cannot justify any valid conclusion as to the shape of the age distribution of the sample. They think this is due to the random fluctuations of the ageing process, in both the *reference population* and the *observed population*. They, therefore, only use this method to provide estimates of the mean of the age distribution, whatever its actual shape. This estimate of the mean they see as sufficiently precise to be accepted with greater confidence. However, where the estimate of age structure is as unlikely as that in the example provided by Masset (1982), cited above, the average age calculated from this structure can hardly be more reliable.

As we have already said, the approximation method was devised empirically, after its supporters had noted that the method presented in the earlier section did not give them satisfactory results, and had no clear mathematical justification. Not only did they doubt its validity for calculating an age structure, but they even thought that starting from a uniform distribution for age was likely to produce an unsatisfactory solution. One of the authors (Bocquet-Appel, 2005) states:

The uniform distribution was deliberately entrenched to eliminate the influence of the reference anthropological sample (Bocquet-Appel and Masset, 1996) but this prior distribution turns out to be a hindrance. It eliminates from the outset a large number of possible archaeological situations that are not represented by a uniform distribution of skeletons (page 279).

We have already shown, in the case of the simple example with two age groups and two stages, that this method covered all possible distributions by age. Similarly, Konigsberg and Frankenberg (1992) state that the result of the iterations is independent of the age structure taken to start from: although one usually takes a uniform structure, this can be replaced by any other without changing the result of the iterations, since any distribution that sums to unity and has no zero cells is acceptable. The uniform distribution inserted in the proposed solution is no justification in our opinion for rejecting the IALK method, despite what Bocquet-Appel has suggested.

12.2.5.1 The IBFP method

We turn to the iterative method he proposes (Bocquet-Appel, 2005, 2008a; Bocquet-Appel et Bacro, 2008) and calls IBPFP (Iterative Bayesian Proportional Fitting Procedure), an inappropriate name since it involves neither the methods usually called IPFP nor Bayesian methods. It is supposed to avoid the use of a uniform *prior* distribution.

Let us first describe the principle of the method. Since starting from a uniform age distribution appears to impose constraints in finding the age structure of a population for which only the stage structure is observed, it would seem to be more useful to start from more general distributions taken from a space of possible demographic models of mortality. We use, therefore, the data from the reference population to calculate the probabilities for the stage indicator given the age. The reason is that from each of the *prior* age distributions it is possible to calculate the stage probabilities. We construct an indicator that indicates the variation between that probability and the one provided by the observed population. In the end we take the age distribution that minimises that indicator.

The problem is that from one publication to another the method used for this purpose leads to different calculations. First, let us examine the presentation given by Bocquet-Appel in *La paléodémographie* (2005).

The algorithm he proposes is based on the probability of belonging to stage *i*, given that one is in age class *j*, f_{ij}, considered to be accurately known for the reference population. He then starts from each of the prior probabilities of belonging to age class *j*, p_j^0, provided by the various types of distribution considered here (Beta, non-central Beta, Weibull and Bi-Weibull). He attempts first to calculate the posterior probability of belonging to age class *j*, where the probability p_i of belonging to stage *i* is known from the observed population. He obtains (page 296)

$$p_j^1 = \sum_{i=1}^{l} \pi_i \frac{f_{ij} p_j^0}{\sum_{j=1}^{c} f_{ij} p_j^0}$$

This estimate turns out to be identical with the first iteration of algorithm [12.3], corresponding to the IALK method, and not that of the approximation method [12.4]. The result is that the subsequent iterations, given in greater detail (Bocquet-Appel, 2005, p.296), if taken far enough, should lead to the same age distribution whatever prior distribution one starts from. But the author ends the iterations when the variation ε between the observed and estimated frequencies by stage falls below the threshold of 10^{-5}, without giving any reason for this choice.

This manner of operating is confirmed by the example of the Loisy-en-Brie population (Bocquet-Appel, 2005, p. 297-298), presented above as Example 3 of the maximum likelihood method, with *prior* age probabilities

(0.10191 0.23279 0.26755 0.22668 0.1361 ;0.03497).

Taking the 10^{-5} threshold, we calculated that the IALK method leads to age structure

(0.212 0.155 0.332 0.201 0.061 0.039),

compared with Bocquet-Appel's estimate (op.cit., top of page 297) rounded to three decimal places

(0.212 0.156 0.332 0.201 0.061 0.038),

which is practically identical.

But in this case, whatever the values of the Beta distribution parameters, the calculated variations are all in principle very close to the 10^{-5} variation, and it is hard to see how the choice of parameters that give the smallest variation ε can lead to a final result. Nor is it clear what age distribution is to be chosen: the one we start from or the one we arrive at when the variation is 10^{-5}, which are quite different from each other.

Furthermore, if the iterations are continued, all these distributions will tend towards the IALK solution given above

(0.220 0.117 0.332 0.196 0.040 0.045).

This is exactly the solution one obtains by continuing the iterations proposed in Bocquet-Appel's article, taking a variation of 10^{-10} after 5,587 iterations.

The result is that this first approach does not seem to lead to a fully satisfactory solution for estimating an age distribution, whatever family of demographic distributions is taken as a *prior*.

12.2.5.2 The *Iterage* program

To understand more clearly the presentation by Bocquet-Appel (2008a) and Bocquet-Appel and Bacro (2008b) in more recent works, it is necessary to refer to the Fortran program Iterage.for, published in August 2007, which can be downloaded from Bocquet-Appel's site: http://www.evolhum.cnrs.fr/bocquet/index.html.

This time, the authors draw 1,000 equally probable samples with replacement by the bootstrap procedure from each of the age groups in the reference population. Unlike in the previous approach, this population is no longer fixed but can vary from one draw to another. Basically, this procedure is introduced so as to be able to estimate the confidence intervals for the age distribution. For each of these reference populations, the authors then use each of the *prior* probabilities by age, calculated from a mortality model including both usual and crisis mortality, to calculate a distance between the breakdown by stage of the observed population and that obtained by calculation, from each prior probability and the stage structure for each age in the reference population.

For that purpose, however, the authors appear to have returned to the method in the earlier article by running iterations, in this case 1,000, and comparing in each iteration the stage structure of the observed population with that provided by the prior probabilities by age. Here we appear once more to come across the disadvantage mentioned above: for a reference population drawn by bootstrap, all the estimates taken from different prior age structures converge on the same solution. However, since this convergence is a slow one, the 1,000 iterations will produce solutions that appear to be different.

On closer inspection, it can be seen that the formulation of these iterations is almost identical to formula [12.3], except that the observed frequency π_i is replaced by its estimate, with the help of *prior* probabilities p_i:

$$p_i = \sum_{j=1}^{c} f_{ij} p_j^0$$

and its first iteration is therefore transformed as

$$p_j^1 = \sum_{i=1}^{l} p_i \frac{f_{ij} p_j^0}{\sum_{j=1}^{c} f_{ij} p_j^0}$$

In this case, it is easy to verify that p_j^1 equals p_j^0 and there is no point in continuing the iterations. This means that, contrary to what one might imagine from the complexity of this subroutine, one simple calculation leads directly to the result p_i. For each prior probability there corresponds one distance between p_i and π_i.

The authors state that this procedure does not permit any valid estimate of the terms in the matrix f_{ij}, which is now random, but can be used to choose for each sample the *prior* probability by age that provides the smallest distance from the stage structure of the observed population. They then calculate the mean of each of these probabilities and estimate a 95% confidence interval with the various bootstrap estimates. However, although the bootstrap technique can be used when the model is properly specified, it is known that no theoretical

result can be used to validate its results when, as here, an empirical model is used with no sufficient specification.

For this program, the authors slightly modified the Loisy-en-Brie data: but it is easy to use the stage data cited in Bocquet-Appel (2005). They also modified the age groups, using seven instead of six. We showed above that if a number of age groups greater than the number of stages is taken, it is impossible to solve the system of equations. The new program makes it possible and the results can still be compared with the previous one by interpolation (failing which, the calculation of the 753 age probability vectors for six instead of seven age groups becomes unnecessarily cumbersome, and the authors do not supply the formulae used). For the Loisy-en-Brie example, the authors' program leads to the following solution:

$$_{10}p_{20} = 0.125,\ _{10}p_{30} = 0.133,\ _{10}p_{40} = 0.172,\ _{10}p_{50} = 0.199,$$

$$_{10}p_{60} = 0.185,\ _{10}p_{70} = 0.125,\ _{10}p_{80+} = 0.061.$$

The program provides the mean chi-square distance between the observed and calculated stage distribution, 0.090. But this is not a chi-square distance, because it is calculated with the formula

$$\chi^2 = \sum_i m_i \left| \frac{\left(\pi_i - \hat{\pi}_i\right)^2}{\hat{\pi}_i} \right|$$

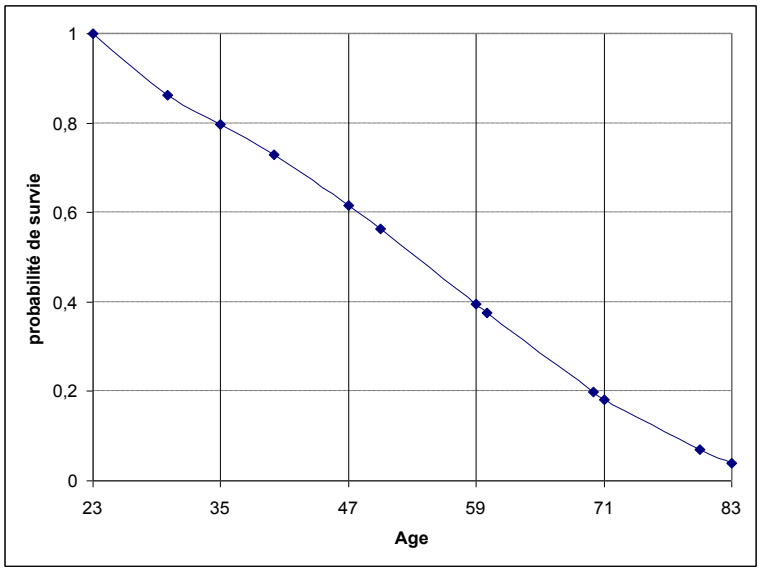
where m_i is the observed number of skeletons at stage i, and π_i and $\hat{\pi}_i$ the frequencies by stage of the observed and estimated population respectively. A standard chi-square distance should include the total number of observed skeletons m and not the number m_i by stage. The effect is that the value calculated by the authors' program underestimates the real chi-square distance. From these proportions of deaths, calculated for the first group over seven years and the following over ten years, it is possible to construct the curve that gives the survival probabilities of this population (Figure 12.2).

From this curve, it is possible to calculate the twelve-year survivors by interpolation (see Figure 12.2) and thus the following structure of age at death:

$$_{12}p_{23} = 0.203,\ _{12}p_{35} = 0.182,\ _{12}p_{47} = 0.220,\ _{12}p_{59} = 0.213,\ _{12}p_{71} = 0.142,\ _{12}p_{83} = 0.040.$$

This structure will be used as a basis for comparison with the Bayesian estimates we provide below.

Figure 12.2 Survival probabilities at Loisy-en-Brie, estimated by Bocquet-Appel and Bacro's *Iterage.for* program, interpolated to obtain survivors every twelve years



However, in the case of the Maubuisson nuns, it is possible to start from the matrix of prior probabilities for normal pre-industrial mortality, known as attritional (20 to 80 and above, in seven ten-year groups), the female Lisbon reference population (see Chapter 4) and the distribution of observed deaths for seven stages, and estimate with the *Iterage.for* program the distribution of death in the seven age groups. This gives

$$_{10}p_{20} = 0.025,\ _{10}p_{30} = 0.036,\ _{10}p_{40} = 0.073,\ _{10}p_{50} = 0.133,$$

$$_{10}p_{60} = 0.210,\ _{10}p_{70} = 0.268,\ _{10}p_{80+} = 0.255.$$

This breakdown will also be used to compare with the Bayesian estimates we make below, particularly since we have an exhaustive estimate of these deaths during the period 1670-1789:

$$_{10}p_{20} = 0.012,\ _{10}p_{30} = 0.025,\ _{10}p_{40} = 0.087,\ _{10}p_{50} = 0.170,$$

$$_{10}p_{60} = 0.289,\ _{10}p_{70} = 0.210,\ _{10}p_{80+} = 0.207.$$

It can already be seen that, although the deaths in the first two ages are reasonably estimated, the quality of the estimates declines sharply for the following age groups.

Although this method does introduce a random element into the reference population, it is still not fully Bayesian, because the observed population is considered here to be non-random. The effect is that the variances calculated by the program must be strongly biased: which is why we have not addressed the point here. In fact, as we shall see below, it is more important to consider the observed frequencies as random than those of the reference population. And by choosing the age structure from a parametric model of mortality, this method, like the previous one, introduces a structure that is not necessarily verified by past populations. If the solution lies outside the list proposed, the authors have no way of verifying that fact.

12.3 Conclusions

After this overview of the methods used by paleodemographers, we can see first that finding a table with observed stage totals where each term comes as close as possible to an initial table

is not the best method. It assumes that the sum of the distances between each term in the reference and observed matrices must be minimised, without allowing for the asymmetric structure of the table, where it is the stage distribution for each given age which must be considered to be invariant.

A much more satisfactory method is to find a table, still with observed stage totals, where the columns come as close as possible to those of an initial table. This complies with the invariance hypothesis that says that for a given stage the age distribution of individuals is independent of the population. However, a large number of conditions that validate this estimate by iteration are generally left unexplained by paleodemographers, who also fail to mention that, even in acceptable conditions, these methods can lead to unacceptable solutions.

The general use of this method, which amounts to assuming that the observed population must follow one of the distributions in a space of mortality models, may be of interest for providing a *prior* distribution for a genuinely Bayesian model. However, the use Bocquet-Appel and Bacro make of it is not strictly speaking a Bayesian generalisation of the second type of model but rather a highly empirical approach that does not take account of the random nature of the reference data. Furthermore, it only holds if the observed population follows one of these distributions

In the following chapter, co-written with Henri Caussinus, we examine in greater detail the second type of model, generalising it as a fully Bayesian model that allows for the uncertainty in all the data we start from.