WORKING PAPERS AND STUDIES

2004 EDITION

Population and social conditions 3/2003/F/n° 26

Methodology for the calculation of Eurostat's demographic indicators

Detailed report by the European Demographic Observatory

G. Calot, J.-P. Sardon - EDO

ISBN 92-894-7076-3
ISSN 1725-065X
Cat. No. KS-CC-04-004-EN-N

3

EUROPEAN COMMISSION

eurostat

THEME 3
Population and social conditions

Europe Direct is a service to help you find answers to your questions about the European Union

New freephone number:
00 800 6 7 8 9 10 11

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (http://europa.eu.int).

Luxembourg: Office for Official Publications of the European Communities, 2004

ISSN 1725-065X
ISBN 92-894-7076-3

Population and social conditions **3/2003/F/n° 26**

Methodology for the calculation of Eurostat's demographic indicators

Detailed report by the European Demographic Observatory

G. Calot, J.-P. Sardon - EDO

Table of contents

Introduction p. 1

Age at the occurrence of an event p. 5

Mean age at the occurrence of an event p. 6

Conversion of the distributions by age of the resident population on a date other than 1 January into distributions on 1 January p. 8

Calculation of annual crude rates p. 9

Order of live births p. 9

Analysis of day data p. 10

Analysis of divorces p. 11

Chapter 1: Operations prior to processing of data in the processing chain p. 12

A) Breakdown by year of age of multiannual data p. 13

I – *Problems affecting the distribution by age of the cohorts of the population* p. 13

1- The distribution of the cohorts of the population is known in a non-annual breakdown of age (most often five-year) p. 13

a) Sprague multipliers p. 13
Sequence of operations p. 14

b) Improving the estimates obtained using Sprague multipliers p. 15
Sequence of operations p. 15

2- Age pyramids available by year of age for the years surrounding those for which only the distribution by age group is available. p. 15
Sequence of operations p. 16

3- Age pyramids available by year of age for the year immediately preceding those for which only the distribution by age group is available. p. 16
Sequence of operations p. 17

4- Breakdown by year of age of the pyramids of which the last group is less than "99 years or over" p. 17
Sequence of operations p. 17

II – Problems affecting the distribution by age of demographic events p. 18

1- The numbers of demographic events according to age are ascertained in a non-annual breakdown of age (in most cases five-year) p. 18

a) Age pyramids available on 1 January by year of age p. 19
Sequence of operations p. 19

b) Age pyramids available on 1 January, only by age group p. 19
Sequence of operations p. 20
"Adjustment" of a Gompertz law p. 20
"Adjustment" of an empirical distribution to the Gompertz law p. 21

B) Other corrections to be made to the basic data p. 23

1- Rectification of non-declared elements p. 23
Sequence of operations p. 23

2- Conversion of age pyramids drawn up on a date other than 1 January into age pyramids on 1 January p. 23
a) Monthly births available p. 23
Sequence of operations p. 24
b) Monthly births not available p. 24

3- Intercensal estimates of the population p. 24
a) No intercensal revision p. 25
Sequence of operations p. 26
b) Revision only of the total cohort of the population p. 26
Sequence of operations p. 27

4- Harmonizing the margins of the different tables concerning the same variable p. 27

C) Estimating missing data p. 28

1- Gap in a chronological series p. 28

a) The distribution of events by age (or by duration) is not available, but the annual total is available p. 28
Sequence of operations p. 28
b) The distribution of events by age (or by duration) is not available and neither is the annual total p. 29
Sequence of operations p. 29
c) The cohort of the population by age (or by duration) is not available, but the annual total is available p. 29
Sequence of operations p. 30
d) The cohort of the population by age (or by duration) is not available and neither is the annual total p. 30
Sequence of operations p. 30

2- Missing data at end of series (with a view to a geographic aggregation) p. 30
a) The cohort of the population by age is not available p. 30
Sequence of operations p. 31
b) The demographic events are not available p. 31
Sequence of operations p. 32

Chapter 2: Construction of tables of event occurrence p. 33

A) The different types of events p. 33

1- Repeatable events p. 33

2- Non-repeatable events p. 34

3- Observation of absolute numbers of events and of cohorts subject to the risk p. 35

B) Repeatable events p. 36

1- Number of person-years of exposure to the risk p. 36
Sequence of operations p. 36

2- Estimate of the rates and construction of tables of occurrence of repeatable events (such as fertility) p. 37

3- Calculation of rates (incidence) p. 44
a) Age in completed years p. 44
Raw method p. 44
Advanced method p. 44
- Constancy of the risk inside the square p. 44

- Taking into account the variability of the risk according to age inside the square p. 50

b) Age reached during the year p. 52
Raw method p. 52
Advanced method p. 53

4- Conversions of events and rates in another Lexis diagram p. 53
Sequence of operations p. 53

C) Non-repeatable events p. 55

1- Estimating probabilities and constructing tables of occurrence of non-repeatable events (such as mortality) p. 55
a) Age in years completed p. 55
Raw method p. 55
Advanced method p. 56
b) Age reached during the year p. 57
Raw method p. 57
Advanced method p. 57
c) Specificity of mortality at 0 and 1 year p. 58
Sequence of operations p. 59

D) Fertility tables by order p. 62

E) Divorce rate table p. 64

Sequence of operations p. 67

Chapter 3: Indicators deduced from the occurrence tables p. 70

1- Derived indicators deduced from the tables p. 70

2- Precautions to be taken when calculating these indicators p. 73

3- Other derived indicators p. 73
a) Crude rates p. 73
b) Mean cohort of generations subject to the risk p. 74
c) Longitudinal recombination of rates and probabilities: longitudinal indicators p. 74
d) The construction of monthly derived indicators p. 75
e) Short-term extrapolation of annual transversal indicators p. 75
Formulation of derived indicators p. 76

Conclusion p. 77

Annex 1: The construction of life tables p. 78

Annex 2: Short-term demographic analysis p.115

Index of indicators p.140

eurostat

Methodology for the calculation of Eurostat's demographic indicators

Detailed report by the EDO

by Gérard Calot and Jean-Paul Sardon

European Demographic Observatory

June 2002

Introduction

The organisation of the collection of demographic information by the National Statistics Offices with a view to the continuous management of an *international* database containing both *basic* and *derived* information (information resulting from the combination of items of basic information) is contingent upon the answers given to a series of preliminary questions.

- What *fields* of analysis are covered by the database?

The fields of analysis covered by the database in respect of each *State*[1] concerned (and apart from international migrations which are not discussed in this document due to the very specific approach adopted), are the cohort and the general structures of the resident population, first marriage, divorce, fertility and mortality.

The degree of demographic detail with which each of these fields is studied is specified when the basic information is issued. However, it is obvious that an international organisation cannot go into very much detail and that in addition the wide differences in national definitions would quickly make this process unrealistic.

The *period* covered by the database will vary from one country to another and from one type of information to another, depending on the range of data available. As much basic information as possible should be combined about the period since 1960. However, older data dating from the Second World War or even the First World War or for that matter from the start of the 20th century may be included in the database where such data exist, as certain analyses, such as those concerning the ageing of the population or the calculation of longitudinal data, require the mobilization of information observed over a long period of time.

- What *definitions* should be adopted with regard to *statistical units* (for example, the definition of *resident* person, *live birth*, *still birth*, etc.) and with regard to *statistical variables* (definition of the *age* of a person at the occurrence of an event, of the *order* of a live birth, etc.)?

[1] We will not discuss here the infra-national databases.

Although, with regard to the statistical *units*, international recommendations have been drawn up by the UN, which are on the whole followed fairly well by European countries[2], there are hardly any statistical *variables*, and the definitions adopted in the different countries can differ considerably. In general, only one definition exists for a given country (for some, this has changed at some point in the past), and it is the basic information that meets this definition that should be collected and inserted in the database and, where appropriate, published with a note specifying the definition adopted. However, at the level of derived information, whenever possible the summarized indicators of *all* the countries should be presented so that they have the *same* definition and are therefore *directly comparable*.

- How *often* should basic information be collected?

  It seems reasonable to collect basic information *annually* on a specific date in the year so that most countries have a batch of *new* annual data in relation to their previous submission and if possible of *final* data (ideally, they should have the *final* data of the *previous* year).

  However, the rate at which annual data are put together differs from country to country. It will therefore be necessary, in particular when establishing basic information or information derived by *geographical aggregation* (for example, for the *entire* European Union[3]), to manage as efficiently as possible the *diversity*, following an annual collection, of the most recent years (*millésimes*) for which a given item of information is available.

- Are only *annual* data collected?

  It would appear judicious, in an *annual* publication, to include certain *monthly* data in order to provide a means of *dating* in more detail than annually any reversals or sudden variations in the most common derived indicators, especially following changes in legislation or particular political or social events.

  In addition, as we shall see later, monthly data over a long period improve the calculation of tables (first marriage, fertility and mortality) and facilitate the conversion of the annual distributions by sex and age of the resident population on a date *other* than 1 January into distributions on 1 January.

  Furthermore, the follow-up of total *monthly* rates provides a means of obtaining, at any time, a *trend projection* of the phenomenon under review, which gives a convenient and immediate means of estimating *missing* data so that when necessary we can estimate the *annual* number of events and the corresponding total rate in the country under consideration, and then carry out any geographical aggregations.

  Finally, the regular publication of *short-term* information, in particular in the form of *graphs* of *monthly* indicators, provides interesting material for the press and the media in

[2] Although certain difficulties remain, in particular with regard to the notion of resident population, especially in the case of countries which have for a long time been *emigration* countries and where there is a population register: the resident population is therefore often the *de jure* population, i.e. the population which is given in the register and therefore not removed from the register after emigration, while it would be desirable for it to be the *habitually resident population*. The important point is not so much the *definition* of the reference population as its *coherence* with the flows observed by the statistical system. However, these flows are difficult to relate to events which occur *outside the national territory*, even if efforts are made to collect data from abroad through consulates.

[3] We refer here to the *geographic aggregation* of data at the level of the European Union, but the problem is the same for any group of states: countries of the Economic and Monetary Union, countries of the European Economic Area, etc.

general on demographic trends in the medium term and thus promotes the database among a wider audience that extends beyond the circle of demographers.

- What data *validation* tools are used *before* the data are entered in the database?

A critical examination of newly received information is essential *before* they are entered in the database, with checks for (logical or calculation) *coherence* of a batch of annual information and checks for the *likelihood* of each item of data. The judgment that must be exercised with regard to the likelihood of the data is generally based on the *chronological regularity* of certain derived indicators and on the need for a satisfactory explanation for any discontinuities (which may result from changes in definition, the taking into account of the results of a new census without having revised the intercensal cohorts by sex and age and also *hardware errors*, in particular errors involving entry or the creation of electronic spreadsheets).

The control *a posteriori* of the quality of the database must also be organized and give rise episodically to comparisons with national publications and a periodic examination of the regularity of different sets of derived indicators.

- How can you ensure that the database managers are informed of *subsequent revisions* that affect the information contained in the database ?

When a national office is called upon to revise its basic data, for example when *final* data are substituted for as yet *provisional* data or when, following the results of a new census, new annual evaluations of the resident population by sex and age are calculated for the entire intercensal period, it is essential that the international organisation managing the database be informed of this revision and that the new data be inserted in the database.

If it is completely out of the question to ask the national offices to check - if only once - the entire content of the database for their country, the checking of a few key annual figures (total population, annual number of events, etc.) will probably be useful in detecting unknown revisions of the database.

- What *methodology* should be used to establish derived information?

We feel that the methodological options concerning the calculation of derived indicators must be guided by four considerations:

International comparability of the results
Statistical quality of the methods used
Simplicity of such methods
Availability and supply, on request, of information presenting such methods

The (even minor) differences between the indicators established and published by the national offices and those resulting from the methodology adopted at an international level pose a delicate problem. In our view, the principle of *comparability* of results - both from *one year to another* for the same country and, *more especially*, from *one country to another* - must, as a last resort, prevail over other considerations and lead national offices to understand and accept that derived information calculated and published by an international organisation may differ *slightly*[4] from the information appearing in the national publications. A note indicating the possibility of such differences may, however, be inserted into the publications.

[4] They would not differ *considerably*, unless there is an error in the implementation of one or other of the methods.

The objectives of comparability of results and of statistical quality overlap to a great extent: the comparability of the results is essential for a good statistical method (absence of bias, correction of disruptive effects, stability and regularity of the results), especially in an international context. However, statistical quality and simplicity do not by any means go hand in hand automatically. Although it is always possible to explain clearly the aims of a method and the reasons why certain elements have been taken into account and to present the properties of the results to be achieved, it would not be realistic to hope to obtain under all circumstances results calculated using a simple formula. Compromises can certainly be sought between statistical quality and simplicity, but this should not be to the detriment of statistical quality.

The methodological problems posed by the creation of an international database do not relate only to the calculation rules of derived indicators. Here are some problems involving the basic information itself:

> How can we estimate the distribution by *year of age* of the resident population belonging to the *terminal* age group provided by the national office (for example, *75 years or over*)? This distribution is necessary for the *aesthetic* construction of the age pyramid of the country, but it is also desirable to construct life tables by age and is essential to aggregate the cohorts geographically by year of age.
>
> How can we estimate a distribution by *single year of age* when, for a given year, only the totals by *five-year age groups* are available (for example, with regard to births according to the age of the mother: in particular Spain prior to 1971)? This distribution is essential for the calculation of the longitudinal indicators, i.e. by year of birth. Similarly, how can we disaggregate by *single* year of age a cohort relating to an age group, for example a *five-year* cohort of resident population?
>
> How can we rectify cases of *non-declared elements* in the statistical distributions using one or more variables?
>
> How can we resolve *non-coherence* between basic information (for example, annual total of the number of *monthly* events differing from the total for *all ages* of the numbers of events by *age* or by age and *birth order*)?
>
> How, more generally, can we make up for the different *lacunae* in the information available for a given country at a given time, and in particular:
>
> > How can we estimate certain basic information for a country which has not yet calculated it, to permit geographical aggregation?
> >
> > How can we estimate the *annual* number of events when we do not yet have the monthly data relating to the *last* months of the year?
> >
> > How can we estimate the total first marriage or fertility rate when we have only the (possibly provisional) *total number* of events of the year and not yet the *distribution by age* of such events?
> >
> > How can we estimate life expectancy (male and female) at birth when we have only the *total number of deaths* for the year and not yet its *distribution by sex and age*?
> >
> > How can we estimate the completed fertility of a generation which has not yet reached 50 years of age in the year in which the most recent data were observed? How can we estimate the proportion of ever-married persons at 50 years of age for a generation which has not yet reached this age in the latest available statistics?

Since the nature of the basic information available for each country and each year can vary from one country to another and even from one year to another for the same country, the system used to process the basic information must provide a means of obtaining derived information which is flexible enough to permit *in all circumstances* the best possible results, taking into account the data which are available when this processing is being carried out.

One should not lose sight of the fact that the management of a database forms a *whole* and that the choices made upstream of the calculation of the derived indicators condition this calculation.

Age at the occurrence of an event

One of the very first problems we come up against with regard to international *comparability* in the field of demography is defining the *age* of a person at the occurrence of an event: age at marriage (of a new spouse), age of the mother at the birth of a child, age at death, age of a marriage at its dissolution through divorce, etc.

Age *in everyday language* is the age *in completed years* at the time of the event, i.e. the age *at the last birthday*. This definition of age is used by certain countries, such as the United Kingdom, to draw up their statistics of marriages, births and divorces. The definition of age adopted by other countries is that which deduces the age from the *year of birth*: difference, for example, as regards fertility, between the year of birth of the child and the year of birth of the mother. This age, which is also referred to as the age *reached* (i.e. during the civil year of the event), is on average half a year *lower* than the age completed.

It is sometimes especially important to specify which definition of age is being used. This is the reason why the fertility of *teenagers* frequently differs by more than a *quarter*, depending on whether it is measured at less than 20 years of age *completed* or at less than 20 years of age *reached*: the international comparisons are completely distorted as this difference is not taken into account.

Certain countries provide data *both* by age completed and by age reached (*triangles* of the Lexis diagram). This is very often the case for European countries with regard to *death* but less frequently as regards marriages, births or divorces, in particular for less recent years.

Consequently, if we are to ensure the international *comparability* of the results, one of the two conventions concerning age must be adopted. The convention with regard to age *reached* is preferable: on the one hand, the longitudinal addition of the rates by age reached provides indicators which relate to a *unique* year of birth while, on the other hand, the *biases* mentioned below are less pronounced as they relate to rates by age reached rather than rates by age completed.

However, the question now arises, for a country which, in a given year, has information *only* by age completed, of how to *convert* this information into age *reached*. This conversion is necessary if we are to compare it with other countries (or to compare it with itself over time in the event that, at some point in the past, it changed its definition of age), but also to geographically aggregate the data of the same year.

The European Demographic Observatory (EDO) has developed a methodology for the construction of tables for first marriage, fertility and mortality, with both definitions of age. In this way, the *two* systems of rates (and probability of dying) - by age completed and by age reached - are calculated, and the absolute numbers of events are estimated using the *triangle* of the Lexis diagram when the basic data are observed *only* by age

completed or *only* by age reached, which permits geographical aggregations using the two definitions.

However, what is at least as important, this methodology provides a means, on the one hand, of *smoothing* the rate curves and, on the other hand and *more especially*, of *correcting* certain rates affected by significant statistical *biases*, which may exceed 10%, due to historical accidents (the two world wars in particular, especially the first). These biases result from the exceptional seasonal pattern of the birth rate during the years marked by these historical accidents. The rates, at *all* ages, of the generations born during those years are biased, in relative value of a quantity which, for the same generation, varies little with the age and the studied phenomenon (first marriage, fertility, mortality). The *longitudinal* indicators of these generations are consequently undermined by bias that the adopted method eliminates satisfactorily. On the other hand, the *transversal* indicators are relatively unaffected: the correction made to a total rate or to a mean transversal age is generally slight.

To apply the method, knowledge of the *monthly* distribution of births over a long period is desirable, but this is information that is available to almost all European countries. Failing this, the distribution, observed in a census, of the population by *year and month* of birth provides the elements required. If even this information is unavailable, the *uniformity* of the seasonal pattern is accepted, which provides a means of improving the estimation of rates in relation to the results of the raw calculation (uncorrected and unsmoothed rates).

The methodology used to calculate the rates is flexible enough to obtain the best possible estimate of each rate, whatever information is available when this calculation is made. Optionally, it also provides the uncorrected and unsmoothed rates.

It must be pointed out that the estimate of the numbers of births in the triangles of the Lexis diagram, obtained by dividing the number observed in the square by 2, results in absolute numbers by parallelograms, then in rates by age reached calculated using the conventional method, of which the irregularity is manifest when they are compared with the rates provided by the EDO method (cf. graph 1 on next page). However, the estimation of the absolute numbers of events in parallelograms is essential to carry out geographical aggregations.

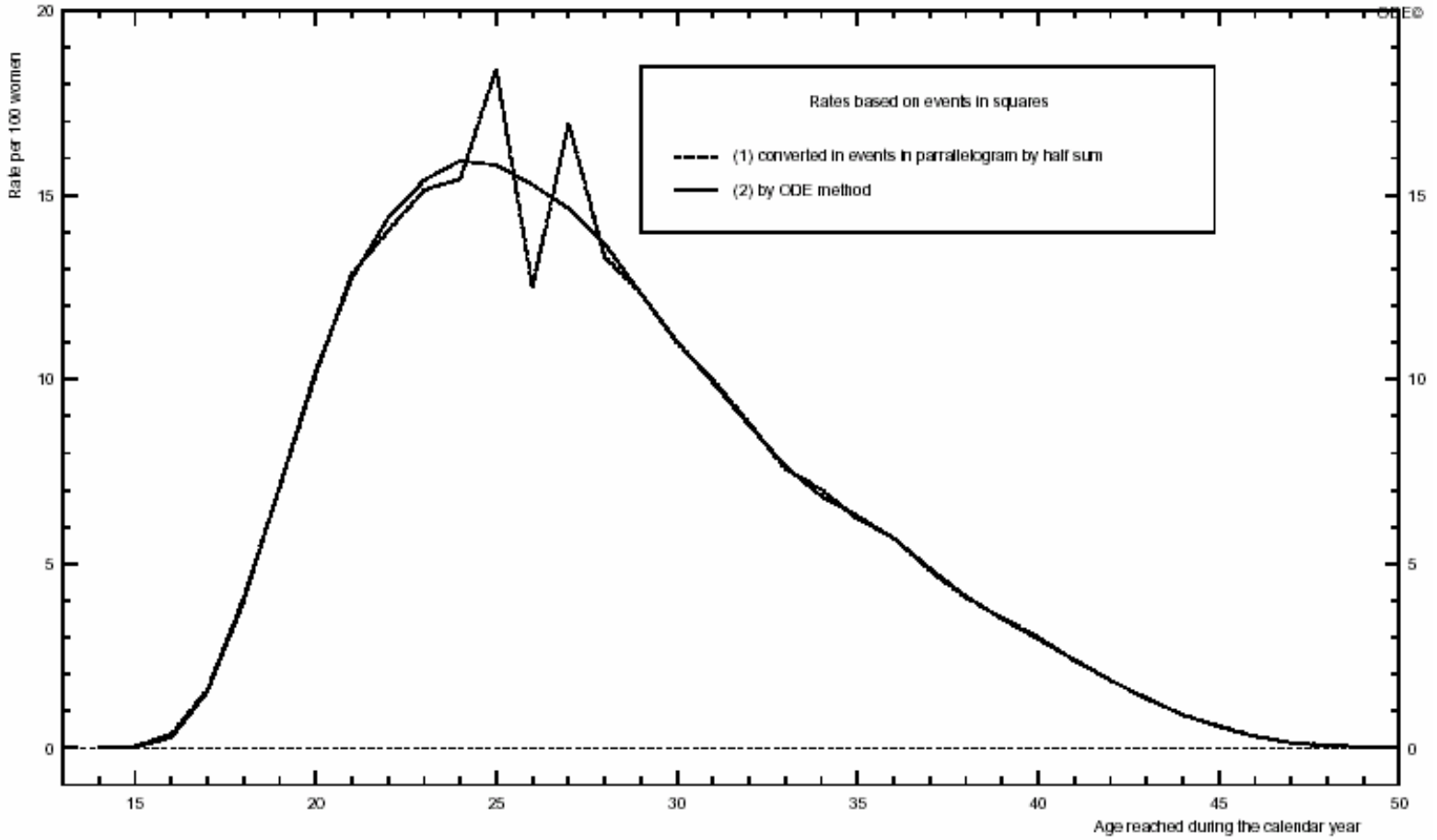
As regards mortality, the document entitled *The construction of life tables* (see annex) presents the methods developed by the EDO to construct the life tables by sex and age.

Mean age at the occurrence of an event

The mean age at the occurrence of an event can be calculated in two different ways: by weighting the ages by the *absolute numbers* of events or by the *rates*.

In the first case, we obtain, for example as regards fertility, the mean age of women who have had a child during the year. This indicator, which matches the traditional definition of a statistical mean (mean age, at childbirth, of the *parturients* of the year), generally corresponds to the information which the medical authorities of the country wish to obtain. However, this mean number is directly affected by the *irregularities* of the pyramid of female ages: if women in the 20-24 age group are particularly *numerous* in the resident population that year (for example, because the baby-boom generations belong to this age group), the mean age will be exceptionally *lower*.

Figure 1. FINLAND, year of observation 1967
Fertility rates, by age of mother reached during the calendar year of birth,
derived from numbers of events observed in squares
(1) converted in numbers of events in parallelogram by half sum of adjacent squares
(2) by ODE method



It follows that the temporal variations in the mean age calculated using absolute numbers as weighting coefficients *simultaneously* reflect two types of variations: those affecting the composition by age of the female population and those concerning the transversal calendar of fertility.

In contrast, the mean age obtained using the *fertility rates* by age as weighting coefficients is not affected by the disruptive phenomenon of the composition by age of the female population and permits comparisons over time and space.

If the differences between the two mean ages were always slight, the *distinction* would not be very significant. However, this is not the case: they can differ by almost one year (whereas a mean age is generally expressed in years to *one* decimal place), thus distorting comparisons.

In an international publication, it is appropriate, for reasons of comparability, to give only mean ages based on the rates.

With regard to first marriage, it is advisable, in addition, to choose a maximum age *below which* the *first marriage* is studied. This maximum age is conventionally the 50th birthday: the frequency of *definitive celibacy* is measured at this age. In spite of the contemporary increase in the age at marriage, we do not think that it is necessary to call into question the choice of this limit for the construction of first marriage tables. The mean age at first marriage, calculated on the basis of rates, must of course relate only to marriages of *single people below* this limit.

Conversion of the distributions by age of the resident population on a date other than 1 January into distributions on 1 January

Most European countries evaluate the cohorts of the resident population by sex and year of age on 1 January of each year. However, certain countries adopt another date: the United Kingdom uses 1 July (*mid-year*), and Ireland uses 15 April. In this case, based on the *monthly* statistics of births over a long period (until the most recent year), the European Demographic Observatory has developed a methodology that allows the cohorts to be evaluated on 1 January.

The systematic conversion of all the evaluations of population by sex and age into evaluations on 1 January provides a means of using the *same* software applications for the construction of tables for all countries and all years and also of geographically aggregating the cohorts by sex and age.

Finally, *whatever* method has been used to calculate the basic information concerning both the cohorts of *residents* by sex and age (evaluations on 1 January or on another date of the year) and the annual *flows* of events by age (first marriages, births and deaths observed by age reached, by age completd or both at the same time), the first marriage, fertility and life tables can be constructed in order to be *directly comparable* from one country to another and from one year to another.

In addition, the different rates are also calculated by year of birth and age completed (*straddling* two years of observation), which provides *cumulations* by generation making it possible to characterize the successive generations by their situation at the time of a *birthday* (for example, proportion of single people who marry before their 25th birthday or between their 25th and 30th birthday, proportion of children who are born between the 25th and 30th birthday of their mother).

Calculation of annual crude rates

The simplest demographic indicators (which are, however, also the roughest) are the *annual crude rates* obtained by calculating the ratio of the total annual number of events observed to the *mean* population of the year. Most European countries use for mean population the half-sum of the cohorts of residents on 1 January of the year and on 1 January of the following year.

However, on the one hand, the population on 1 January of the following year may differ from that on 31 December of the year because the national office has decided to make a *statistical adjustment* to palliate a calculation inconsistency between the variation in the total population, the balance of the natural change (births-deaths balance) and the balance of international migrations[5]. In this case, it seems desirable, in an international publication giving the elements of the population change in absolute numbers and in the form of crude rates (crude rate of variation in the total population, crude rate of natural increase, relative net migration), to use for mean population in the calculation of the latter half-sum of the cohorts on 1 January and on 31 December and to indicate in a note the value of the statistical adjustment.

Furthermore, certain countries that have a population register draw up *monthly* or *quarterly* series of the resident population and use for mean population of the year the arithmetic mean of the monthly or quarterly values. The resulting crude rates, usually expressed in *per thousand* to one decimal place, generally hardly differ except in terms of the effects of rounding of those obtained by reference to the half-sum of the cohorts at both ends of the year. It therefore appears desirable that the crude rates deduced from an international database be calculated *uniformly* for *all* countries using the latter method.

Order of live births

The *order* of a live birth is a variable of great interest for the study of fertility. However, two different definitions of the order coexist in Europe: the biological order, defined in relation to the genesic history of the mother which applies to *all* live births, and the order *in the current marriage* which applies only to births *within the marriage*[6].

At times when births *outside marriage* and divorces and widowhoods, and therefore *remarriages*, are rare, the two definitions are fairly similar. However, when any of these are frequent, in particular during the contemporary period, the order in the current marriage loses much of its *relevance*. However, quite a number of European countries - in particular Germany, France and the United Kingdom - still produce civil status statistics only according to the definition of the order in the current marriage. Other countries have produced annual data by biological order only since relatively recently (Sweden since 1974, Austria since 1984).

Therefore the question arises as to how we can ascertain whether, for countries which do not have annual data by biological order, it would be desirable to show in the international publications, *in addition to* – and possibly even *instead of* – traditional annual series based on civil status statistics, some data (which could not really be

[5] This situation occurs when, following a difference in the degree of completeness of two successive censuses, the estimate of the intercensal variation of the total population is not equal to the sum of the balance of the natural change and of the balance of the international migrations.

[6] It should be remembered that certain countries determine the birth order by adding, as appropriate, one unit to the number of *previous* births, but including in this number stillborn or adopted children. Moreover, it may be that the statistics of births by order are not limited to live births, but also cover stillbirths or all childbirths.

annual) obtained according to the biological order through *surveys* or thanks to some reliable method of estimation. This observation applies to the *transversal* indicators (by year or period of observation) and to the *longitudinal* indicators (by year or period of birth of generations), in particular for the proportion of women *without any children* (measurement of infertility) in the generations. The same applies more generally to the breakdown of the completd fertility according to the final number of children.

This comment ties in with the next aspect, which relates to another weakness of civil status statistics. The events counted by these statistics are generally limited to those occurring on the *territory* of the country considered. However, international migrations can cause more or less marked discrepancies between the *flow* data recorded by the civil status and the *stock* data obtained during censuses. For example, with regard to first marriage, certain immigrants who entered the country as single persons may return to their country of origin to get married; in this case, their marriage escapes the statistics of the host country but they are counted as married if a census has just been carried out there. For this reason, it is desirable, at least for the last intercensal period, to compare systematically the increase in the proportion of ever-married persons per generation obtained through the cumulation of the civil status statistics during the intercensal period and that obtained by calculating the difference between the statistics of the two censuses. This comparison also allows any errors in the database to be detected.

This same type of comparison between flows and stocks can be carried out with regard to the distribution of women according to the number of children already born[7] or the proportion of persons born outside the country according to sex (test of coherence between the intercensal variation of this proportion according to the generation and the net migrations by sex and age of the intercensal period).

For the countries that have, for quite a long period of observation, annual information on live births according to the *biological* order, it is suggested that the *parity progression ratios* be calculated, i.e. the frequencies of passage from r to $r + 1$ children according to the age of the mother, as well as their aggregates *all ages combined* (transversal and longitudinal). Similarly, as regards first marriage (male and female), it is suggested that we calculate not only the rates analogous to the fertility rates (rates also referred to as the *occurrence-exposure* rate in the literature), but also the *first marriage* probabilities (or *incidence* rate) and their transversal summaries (total rates and mean ages based on the probabilities).

Analysis of daily data

Although not essential, the collection of daily data with regard to births and especially marriages is desirable to refine the correction of seasonal variations. While marriages in a week are increasingly concentrated in most European countries on a given day (Saturday in certain countries, Friday in others), a crude number of marriages relating to the same month of two different years will not have the same significance depending on whether the corresponding month has four or five Saturdays (or Fridays). These daily data are frequently calculated by the national offices for statistical checks, and are therefore often available. Also, the day data as regards births provide interesting indications on medical practices.

[7] Provided that the civil status statistics are calculated according to biological rank.

Analysis of divorces

The divorce rate is analyzed, as is the case for fertility, according to the age of the mother by means of the *rates* based on the age of the marriage. However, although the national offices estimate the female cohorts by age on 1 January of each year, taking into account both mortality and migrations, they do not keep an analogous up-to-date record of subsisting marriages on each 1 January, considering both dissolutions of marriages (by widowhood or divorce) and migrations. Also, the only divorce rates that can be calculated give the ratio of the numbers of divorces observed with a given age to the *initial* number of weddings in the country. These rates are affected by the disruptive effect of widowhoods and especially of migrations.

In addition, the definition of divorce varies from country to country. Some countries calculate their annual statistics on the basis of the number of divorce rulings *pronounced* by the courts while others calculate their statistics based on divorces *transcribed* in registry of births, marriages and deaths. However, certain divorces are never transcribed (for example, because neither of the ex-spouses remarry), while a divorce pronounced in a given year, if it is transcribed, may in fact be transcribed several years after the ruling.

The conversion of the divorce rates by age of the marriage reached or completed may be carried out with the same concern for statistical quality only as regards fertility, first marriage or mortality, especially as the available information is sometimes only by five-year age groups. However, the transversal divorce rates can be added longitudinally to evaluate the intensity of the divorce in successive marriage cohorts.

Chapter 1

Operations prior to processing of data in the processing chain

The processing of data received from national statistics institutes consists, if no anomalies are detected in such information, in the application of different statistical methods, at the end of which files of results are created which we shall refer to as *derived files*. If no major anomalies are detected in the information received, the derived files must be able to be created completely automatically. However, different kinds of anomalies may be encountered.

- Some anomalies will be resolved by the processing chain itself (for example, frequent anomalies or anomalies of which the correction does not involve a complex calculation). This first category includes those relating to:
  1. the definition of age (age reached and age completed),
  2. non-declared elements,
  3. differences in the total numbers of events according to the classification criteria (for example, live births according to the age of the mother, the birth order or the month of occurrence).

- Other anomalies, however, will result in rejection by the automatic processing chain, requiring the *prior* correction of data received from the national office. This prior correction may itself be carried out:
  1. manually
  2. or automatically, using an *ad hoc* program, possibly after consultation with the national office. Prior to submitting the official data of a country for processing which will necessarily modify them automatically, it is generally desirable to contact the persons responsible for the statistics of that country to explain to them the modifications envisaged and, where possible, to secure their agreement. It is advisable, in particular, to reach an understanding with the country when there are several possible options.

  The anomalies of this second category, which will in most cases relate to *old* periods, will be resolved taking into account the information available when the processing is carried out. In general, this will consist of a specific processing technique which is carried out once only based on the data received from the national office.

In this first chapter, we will discuss anomalies of the second type which require the application of preliminary processes to the data received from the national offices, so that they meet the constraints of the automated chain.

We will discuss three specific cases in sequence:

- Breakdown by year of age of the data regrouped;
- Other corrections to be made to the database;
- Estimation of missing data.

A) Breakdown by year of age of multiannual data

I - Problems affecting the distribution by age of the cohorts of the population

1- The distribution of the cohorts of the population is known in a non-annual breakdown of age (most often five-year)

The annual estimations of *population by sex and age* may be available, not in an *annual breakdown* that covers *all* ages, but in a *less detailed* breakdown. In the rest of this chapter, we will use the term *five-year* to indicate a *non-annual* breakdown over the range of ages studied (it may be annual in certain age zones and multiannual, possibly with class amplitudes that *vary* from one class to another, at other ages).

This non-availability of the distribution by sex and year of age of the population prevents the direct calculation of the demographic rates relating to only one cohort, which considerably complicates the longitudinal recombination of these rates and thus the updating of the indices relating to particular cohorts or generations.

It also prevents the aggregation of these data with those corresponding to a more refined breakdown, unless a similar regrouping of these data is carried out.

There are several options:

1. We can decide not to take into consideration those years for which the breakdown is not annual;
2. We can use the existing breakdown and adapt it to the breakdown of events to make it compatible with that of the population, which allows us to make a calculation of the transversal indicators and rates, but does not permit longitudinal transposition, unless we consider that the rates are constant over the entire age group;
3. Using the existing breakdown, we can calculate rates by age group and then annualize the curve of cumulated rates, based on the adjustment to the Gompertz law;
4. We can break down the cohorts by five-year age groups into cohorts by year of age.

If the last option is chosen, there are several methods which provide a means of carrying out a breakdown of the five-year age groups of any *age pyramid*. They are based on all the regularities observed in the transition from one age to the next and can apply in all circumstances since they require knowledge only of the pyramid to be annualized. The best-known of these methods uses the *Sprague multipliers*[8].

a) Sprague multipliers

This method is applied as follows:

Let us take a five-year group of cohorts N_0 that we wish to break down into five cohorts by age:

$$n_1 + n_2 + n_3 + n_4 + n_5 = N_0.$$

[8] Cf. Thomas Bond Sprague, "Explanation of a New Formula for Interpolation", *Journal of the Institute of Actuaries*, 22:270, 1880-1881.

We must first identify the first two and the last two five-year groups that require specific processing, to which we shall return below.

For the other groups, the *intemediate table* of Sprague multipliers is used (of which the totals in columns are equal to 0 or 1 and those in rows to 0.2):

	N_{-2}		N_{-1}		N_0		N_{+1}		N_{+2}		Total	
n_1	-	0.0128	+	0.0848	+	0.1504	-	0.0240	+	0.0016		*0.2000*
n_2	-	0.0016	+	0.0144	+	0.2224	-	0.0416	+	0.0064		*0.2000*
n_3	+	0.0064	-	0.0336	+	0.2544	-	0.0336	+	0.0064		*0.2000*
n_4	+	0.0064	-	0.0416	+	0.2224	+	0.0144	-	0.0016		*0.2000*
n_5	+	0.0016	-	0.0240	+	0.1504	+	0.0848	-	0.0128		*0.2000*
Total		*0.0000*		*0.0000*		*1.0000*		*0.0000*		*0.0000*		*1.0000*

Sequence of operations

To estimate, for example, the cohort at 27 years, which corresponds to n_3 in the five-year age group 25-29, we have:

$$N_{-2} = N_{-15\text{-}19}$$
$$N_{-1} = N_{-20\text{-}24}$$
$$N_0 = N_{-25\text{-}29}$$
$$N_{+1} = N_{-30\text{-}34}$$
$$N_{+2} = N_{-35\text{-}39}$$

Using the coefficients of row n_3, we write:

$$n_3 = +0{,}0064\ N_{-15\text{-}19} - 0{,}0336\ N_{-20\text{-}24} + 0{,}2544\ N_{-25\text{-}29} - 0{,}0336\ N_{-25\text{-}29} + 0{,}0064\ N_{-35\text{-}39}$$

which gives the cohort of individuals who are 27 years old.

For the extreme groups, we use coefficients which involve only four five-year groups.

For the extreme groups, we use coefficients which involve only four five-year groups.

Table for the *first* five-year group

	N_0		N_{+1}		N_{+2}		N_{+3}		Total	
n_1	+	0.3616	-	0.2768	+	0.1488	-	0.0336		0.2000
n_2	+	0.2640	-	0.0960	+	0.0400	-	0.0080		0.2000
n_3	+	0.1840	+	0.0400	-	0.0320	+	0.0080		0.2000
n_4	+	0.1200	+	0.1360	-	0.0720	+	0.0160		0.2000
n_5	+	0.0704	+	0.1968	-	0.0848	+	0.0176		0.2000
Total		*1.0000*		*0.0000*		*0.0000*		*0.0000*		*1.0000*

Table for the *second* five-year group

	N_{-1}		N_0		N_{+1}		N_{+2}		Total	
n_1	+	0.0336	+	0.2272	-	0.0752	+	0.0144		0.2000
n_2	+	0.0080	+	0.2320	-	0.0480	+	0.0080		0.2000
n_3	-	0.0080	+	0.2160	-	0.0080	+	0.0000		0.2000
n_4	-	0.0160	+	0.1840	+	0.0400	-	0.0080		0.2000
n_5	-	0.0176	+	0.1408	+	0.0912	-	0.0144		0.2000
Total		*0.0000*		*1.0000*		*0.0000*		*0.0000*		*1.0000*

Table for the *penultimate* five-year group

	N_{-2}		N_{-1}		N_0		N_{+1}		Total	
n_1	-	0.0144	+	0.0912	+	0.1408	-	0.0176		0.2000
n_2	-	0.0080	+	0.0400	+	0.1840	-	0.0160		0.2000
n_3	+	0.0000	-	0.0080	+	0.2160	-	0.0080		0.2000
n_4	+	0.0080	-	0.0480	+	0.2320	+	0.0080		0.2000
n_5	+	0.0144	-	0.0752	+	0.2272	+	0.0336		0.2000
Total		*0.0000*		*0.0000*		*0.0000*		*0.0000*		*1.0000*

Table for the *last* five-year group

	N_{-3}		N_{-2}		N_{-1}		N_0		Total	
n_1	+	0.0176	-	0.0848	+	0.1968	+	0.0704		0.2000
n_2	+	0.0160	-	0.0720	+	0.1360	+	0.1200		0.2000
n_3	+	0.0080	-	0.0320	+	0.0400	+	0.1840		0.2000
n_4	-	0.0080	+	0.0400	-	0.0960	+	0.2640		0.2000
n_5	-	0.0336	+	0.1488	-	0.2768	+	0.3616		0.2000
Total		*0.0000*		*0.0000*		*0.0000*		*0.0000*		*1.0000*

b) Improving the estimates obtained using Sprague multipliers

This method is limited in the fact that it does not mobilize all the available information. However, to obtain the best possible estimate of the distribution of the population by year of age, it is advisable to take account of as much information as possible. To this end, it may be necessary to identify different situations according to the type of data on which these estimates may be based. Thus, in a second stage, if the numbers of annual live births are available over a long period or if a detailled breakdown by age and sex at a specific date (before and/or after the date of estimation) is available, it is possible, *if desired*, to calculate the apparent survival coefficients, and to subsequently smooth them to determine, based on these smoothed coefficients, a distribution by age, which is probably closer to the actual distribution.

Sequence of operations

If SP_x is the cohort estimated at age *x* based on Sprague multipliers,

and $(1-_x qa_0) = \frac{SP_x}{N_{n-x}}$ is the apparent survival coefficient from birth to age *x*,

then:

1. calculate the sequence of ratios $\frac{SP_x}{N_{n-x}}$,
2. smooth the series of ratios obtained by mobile means to obtain $(1-_x qa_0)''$,
3. obtain the improved estimates $SP_x'' = N_{n-x} * (1-_x qa_0)''$.

2 - Age pyramids available by year of age for the years surrounding those for which only the distribution by age group is available.

When we have the distribution by year of age of the population for the years which surround those for which only the distribution by *age group* is available, the procedure to be implemented is simple and consists of two stages:

- First, a linear interpolation is made between the cohorts of the same generation between *n*-1 and *n*+1;

- Then, by proportional correction, the five-year sums of the interpolated cohorts are made to coincide with the corresponding five-year cohort observed.

With this procedure for estimating the cohorts by generation, the only difficulty lies in determining the cohorts of the first and last age: zero and 99 years and over.

The cohort at 0 years will be estimated by applying to the number of births of year *n*-1 an *apparent*[9] survival coefficient at 0 years completed on 1 January of year *n*. That of the open terminal group (99 years and over) will be obtained by applying to each of the ages (98 and 99 and over) a survival probability of 1 additional year, then by calculating the sum of the two numbers thus obtained.

This procedure is identical to that used when the estimate by age and sex of the population for a given year is not available[10].

Sequence of operations

1- Linear interpolation: $P_n^g = \frac{P_{n-1}^g + P_{n+1}^g}{2}$,

2- Proportional correction: $\sum_g^{g+a} P_n^g = P_n^{g,g+a}$

$$P_n''^g = P_n^g * \frac{P_n^{g,g+a}}{\sum_g^{g+a} P_n^g}$$

3- Cohort at 0 years: $P_n^0 = N_{n-1} * (1 - ka_0)$, where ka_0 = apparent *partial* survival coefficient

4- Cohort at 99 years and over: $P_n^{99+} = \sum_{98}^{W} \left(P_{n-1}^{98} * (1 - q_{98})\right) + \left(P_{n-1}^{99} * (1 - q_{99})\right) + \left(P_{n-1}^{100} * (1 - q_{100}\right) + ...$

$$P_n^{99+} \approx \sum_{98}^{W} \left(P_{n-1}^{98} * (1 - q_{98})\right) + \left(P_{n-1}^{99+} * (1 - q_{99})\right)$$

3 - Age pyramids available by year of age for the year immediately preceding those for which only the distribution by age group is available.

When we have *only* the distribution by age of the population of the year immediately preceding those for which only the distribution by age group is available, the method by interpolation described above cannot be used. Thus, for each of the successive years we must choose apparent *perspective* survival coefficients between 1 January and 31 December of the year, apply them to each age and then use the procedure described in the previous paragraph to estimate the cohorts aged 0 and 99 and over on 1 January of year *n*.

Each five-year sum of estimated cohorts is then, through proportional correction, made to coincide with the five-year cohort observed.

Subsequently, the distribution, by year of age, of the resident population of each of the years for which only the five-year distribution is available is estimated through successive iterations of this procedure.

[9] Since it involves the net migration.

[10] We will present this in more detail when we discuss missing data in chapter I C, page 28.

Sequence of operations

1- Choice of apparent survival coefficients: $(1-qa_x)$,

2- Calculation of the cohorts of each age in year *n*: $P_n^x = P_{n-1}^{x-1} * (1-qa_{x-1})$,

3- Cohort at 0 years: $P_n^0 = N_{n-1} * (1-ka_0)$,

4- Cohort at 99 years and over: $P_n^{99+} = \sum_{98}^{W} \left(P_{n-1}^{98} * (1-q_{98})\right) + \left(P_{n-1}^{99} * (1-q_{99})\right) + \left(P_{n-1}^{100} * (1-q_{100}\right) + ...$

$$P_n^{99+} \approx \sum_{98}^{W} \left(P_{n-1}^{98} * (1-q_{98})\right) + \left(P_{n-1}^{99+} * (1-q_{99})\right),$$

5- Proportional correction: $\sum_{x}^{x+a} P_n^x = P_n^{x,x+a}$

$$P_n^{\prime\prime x} = P_n^x * \frac{P_n^{x,x+a}}{\sum_{x}^{x+a} P_n^x}$$

4 - Breakdown by year of age of the pyramids of which the last group is less than "99 years or over"

When the pyramids by *year of age* end with an open group of which the lower limit is *less than* 99 years (for example, as was the case for a long time in many European countries, terminal age group: 85 years or over), it is advisable to break down by year of age the terminal group to permit an aesthetic graphic representation of the age population, but especially to construct the life table.

In this type of situation, the ideal solution is to obtain from the national statistics office the distribution according to the desired breakdown if this exists or if it can be estimated by the office. Otherwise, it is the responsibility of Eurostat to carry out the breakdown of this open group.

To extend the pyramid, a life table is used which is adapted to the conditions of the country in question, generally a table from a recent year for this country or a neighbouring country.

This life table can be used in two different ways, according to the type of population associated with the life table chosen:

- The first way presupposes that the population in question is the *stationary* population associated with the life table, but with a discrepancy that varies linearly with age.
- The second way presupposes that the population is a *stable* population associated with the table used, whereby the growth rate of this population is to be determined.

The method that results in the estimated cohorts varying the most regularly according to age is adopted.

Sequence of operations

1- Stationary population

- estimation of each of the cohorts of the last 10 ages observed and 99 years, applying to the cohorts of the stationary population the rule of proportionality

between the total of these cohorts in the stationary population and in the population observed;

- calculation of the ratio between the cohorts observed and those, with the same total, estimated above based on the stationary population;
- estimate of the linear discrepancy;
- calculation of the previous adjusted ratio of the linear discrepancy;
- adjusted estimate of the cohorts by application of the adjusted ratio above;
- final estimate, applying to the adjusted estimate the rule of proportionality between the total of the open group observed and that of the stationary population.

2- Stable population

- estimation of the growth rate based on the ratio of the cohorts of each of the last 10 ages, observed and deduced from the stationary population;
- determination of the cohorts of the stable population based on the previously determined growth rate;
- estimation of each of the cohorts between the last age observed and 99 years, applying to each of these cohorts the rule of proportionality between the total of the open group observed and that of the stable population.

II) Problems affecting the distribution by age of demographic events

1 - The numbers of demographic events according to age are ascertained in a non-annual breakdown of age (in most cases five-year)

The transversal statistics of *flows of events by age* (births of all rows or of one specific row, first marriages of a specified sex, death of a specified sex, divorces) may be available, not in an *annual breakdown* covering *all* ages, but in a *less detailed* breakdown. As above, we will use the term *five-year* to indicate a *non-annual* breakdown over the range of ages studied (it may be annual in certain age zones and multiannual, possibly with class amplitudes that *vary* from one class to another, with other ages).

Two cases can be identified:

- *age pyramids* are available on 1 January of each civil year *by year of age* over the *entire range* of the ages concerned by the flow of events considered;
- *age pyramids* are available on 1 January of each civil year only *by age groups.*

The absolute numbers *not declared* must of course be rectified *before* any other calculation is made.

a) Age pyramids available on 1 January by year of age

The *absolute numbers* of *five-year* events and the *five-year rates* must be ranked in specific places defined by the domain (for example, live births, live births of biological order equal to 2, births outside marriage, male first marriages, female deaths, or divorces) and by the figure of the Lexis diagram to which these aggregates refer (CARRE, PV or PH).

Five-year data is converted into *annual* data by "adjusting" the *five-year rates* using a Gompertz law[11] of which the cumulative function coincides with that of the five-year rates observed.

The *estimated annual* rates are those of the "adjusted" Gompertz law. From these *estimated annual* rates, we can deduce, using the cohorts of population by age, the *estimated absolute annual numbers* of events.

These *estimated* absolute numbers by year of age are then processed by the automatic chain as if they were *observed* numbers.

Sequence of operations

1- Entry of the cohorts observed, as if they were by year of age over the *entire* range of ages, allocating the cohorts of one five-year group to the *central* age of this group (the 35-39 age group is allocated to 37 years of age). The non-declared elements are, as usual, provisionally allocated to the maximum age plus one unit.

2- Automatic rectification of the non-declared elements, keeping a copy of the initial file which included the non-declared elements.

3- Creation of the file of the *absolute numbers by a rigorous five-year age group*.

4- Creation of the file of the *raw rates by a rigorous five-year age group*: $t_n^{x,x+5} = \frac{E_n^{x,x+5}}{\sum_{x}^{x+5} \overline{P_n^x}}$

5- "Adjustment" of the Gompertz law and creation of the file of the *rates "adjusted" by year of age* $t_n''^x$.

6- Creation of *two* files: first, a file that contains the *absolute numbers "adjusted" by year of age* and, second, a file that contains the *absolute numbers of year of age finally selected*: $E_n''^x = \overline{P_n^x} * t_n''^x$

b) Age pyramids available on 1 January, only by age group

When the age pyramids are not available by *year of age* but only by *age group*, in a breakdown of the age *coherent* with that of the absolute numbers of events by age, we start by establishing the files of five-year rates using an *ad hoc* program. Their annualization, as mentioned above, is calculated using an "adjustment" of the Gompertz law. However, it is not possible to deduce from the estimated annual rates the estimated absolute annual numbers.

[11] The adjustment procedure is described a little further on in this document.

If the age pyramids available are *rigorously* five-year and the absolute numbers of events are available in *squares* by *rigorous* five-year age groups, an *ad hoc* program is necessary.

The annual rates thus estimated can be used within the framework of the longitudinal recombination to supplement the information available for each of the cohorts and to make up for the absence, whether partial or total, of data by year of birth or year of age.

Sequence of operations

1 and 2- Idem page 19

3- Creation of the file of the *absolute numbers by a rigorous five-year age group*.

4- Creation of the file of the *raw rates by a rigorous five-year age group*: $t_n^{x,x+5} = \frac{E_n^{x,x+5}}{\sum_x^{x+5} \overline{P}_n^x}$

5- "Adjustment" of the Gompertz law and creation of the file of the *rates "adjusted" by year of age* $t_n''^{x,x+5}$.

"Adjustment" of a Gompertz law

Gompertz's law is used in terms of the *cumulative function* of a statistical distribution as follows:

The *continuous* statistical variable *X* follows the Gompertz law if its cumulative function *F*(*x*), the proportion of individuals of which the character *X* is less than *x*, matches the formula:

$$F(x) = \exp\left\{-\exp\left[P_k\left(\frac{x-b}{a}\right)\right]\right\}, \text{ i.e. } \text{Log}\{-\text{Log}[F(x)]\} = P_k\left(\frac{x-b}{a}\right),$$

expressions where *a* and *b* are *constants* and where $P_k(x)$ is a polynome of degree *k* in •.

For this to be a genuinely cumulative function, i.e. an *increasing monotone* function varying from 0 to 1 when *x* varies from $-\infty$ to $+\infty$, it is particularly essential that degree *k* of the polynome should be *odd* and that the coefficient a_k of degree *k* is *negative*. However, the property of increasing monotone function is only really necessary in the variation interval of *X*: outside this interval, it does not matter whether or not the variations of *F* are monotone and whether or not the values of *F* are between 0 and 1.

The *quantile* of order a of this distribution is equal to:

$$X_a = ax + b$$

where • is the root, which must be *unique*, of the equation:

$$P_k(x) = \text{Log}[-\text{Log}(a)]$$

It should be noted that the parameters *a* and *b* are *false* parameters: if a Gompertz law is applied, whereby the values of *a*, *b*, a_s, *s* = 0, 1, …, *k* are fixed, this same law can be applied with other *arbitrary* values of coefficients *a* and *b* (with which are associated corresponding values of coefficients a_s). If $\text{Log}\{-\text{Log}[F(x)]\}$ is a polynome of degree *k* in $\frac{x-b}{a}$, it is also a polynome of degree *k* in $\frac{x-\hat{b}}{\hat{a}}$, whatever the values of $\hat{a}$ and $\hat{b}$. The *multiplicity* of the Gompertz laws of which polynome *P* is degree *k* is thus order *k*+1.

[12] We use the term *adjusted*, although it is not actually an adjustment of the cumulative curve but the determination of Gompertz's cumulative curve which goes *exactly* through a chosen given *sub-universe* of empirical points. We therefore refer to adjustment and non-adjustment.

"Adjustment" of an empirical distribution to the Gompertz law

Let us assume an empirical distribution of which the cumulative function *F* is given in the form of values of *F* corresponding to *r*+1 upper class limits: related to the upper class limit x_j is the value F_j ($0<F_j<1$) of the cumulative function, j = 1, 2, …, *r*+1. If the values of *a* and *b*, degree *k* of the Gompertz polynome and the *k*+1 couples (x_i, F_i) are fixed, i = 1, 2, …, *k*+1 for which we force the adjusted curve to pass, polynome *P* of degree *k* is completely determined.

Its coefficients a_s, s = 0, 1, …, *k* are the roots of the system of *k*+1 linear equations with *k*+1 unknown:

$$\sum_{s=0}^{s=k} a_s \left(\frac{x_i - b}{a} \right)^s = \quad \mathrm{Log}[-\mathrm{Log}(F_i)]$$

In these conditions, the *adjusted*[12] cumulative curve goes exactly through the *k*+1 points (x_i, F_i) *selected* to determine the coefficients a_s, but it is in no way guaranteed that the values which are deduced therefrom for these coefficients make the function $\exp\left\{-\exp\left[P_k\left(\frac{x-b}{a}\right)\right]\right\}$ an increasing monotone function with values between 0 and 1. In particular, it may well be that the adjusted frequency of a given class, let us say class ($x_u \le x \le x_{u+1}$), i.e. the quantity:

$$\exp\left\{-\exp\left[P_k\left(\frac{x_{u+1}-b}{a}\right)\right]\right\} - \exp\left\{-\exp\left[P_k\left(\frac{x_u-b}{a}\right)\right]\right\},$$

is not between 0 and 1, although, for *all* the classes resulting from the selection made to determine the coefficients a_s, the adjusted frequencies are *exactly* equal to the empirical frequencies.

Two cases can be identified:

a) we fix degree *k* of the polynome to be adjusted, but *a priori* we do not fix the *k*+1 values x_i through which we force the curve of cumulative frequencies to pass

b) we fix the *k*+1 values x_i through which we force the curve of cumulative frequencies to pass.

In case b), the cumulated frequencies F_i associated with the values x_i are the empirical cumulative frequencies deduced from the five-year frequencies (not taking into account any very small frequency classes, such as Less than 15 years of age or 45 years of age or over) and the degree of polynome *k* is, for example, taken to be equal to 5 if the $k + 1 = 6$ classes selected are Less than 20 years of age to 40 years of age or over

In case a), to limit the risk of certain "adjusted" cumulative frequencies being negative or greater than 1, it is advisable to select the *k*+1 points that are used to make adjustments as follows:

- First value selected so that *F* is quite close to 0 but not too close, let us say that *F* is around 1%
- Last value selected so that *F* is quite close to 1 but not too close, let us say that *F* is around 99%
- Other values x_i selected so that the differences $F_i - F_{i-1}$ are quite close to one another, but smaller for $i = 2$ and $i = k$

In practice, if a polynome of degree *k* is adjusted, we will endeavour to select the *distinct* *k*+1 values x_i on the basis of the empirical cumulated frequencies F_j so that we are as close as possible to the situation where:

- for i = 1: F_1 = 1%
- for i = 2: $F_2 = (1 + \frac{49}{k-1})\%$
- for any i from 2 to k: $F_i = (1 + \frac{49\,(2i-3)}{k-1})\%$
- for $i = k$: $F_k = (1 + \frac{49\,(2k-3)}{k-1})\%$

- for $i = k+1$: $F_{k+1} = 99\%$

Thus, when we want to adjust a polynome of degree 6, we select 7 class limits x_i with approximately:

$$F_1 = 1\%,\ F_2 = 11\%,\ F_3 = 30\%,\ F_4 = 50\%,\ F_5 = 70\%,\ F_6 = 89\%,\ F_7 = 99\%$$

B) Other corrections to be made to the basic data

1- Rectification of non-declared elements

One of the difficulties encountered in the international comparisons lies in the treatment of non-declared elements in the statistics tables. Some national offices publish tables containing specific rows and columns for non-declared elements, while others carry out this rectification prior to publication, without indicating the possible existence of non-declared elements or the procedures used.

In a database that supplies an indicator calculation system, *all* the tables that contain non-declared elements must be rectified. However, a copy of each of the tables must be kept in their original form so as to provide a means of tracing any errors at a later date.

The rectification procedure generally selected consists in distributing the non-declared elements according to the elements declared. It is in fact difficult to define for all the countries and phenomena studied simple procedures based on selection hypotheses for populations on which information is missing.

For this method - which is certainly raw but which has the advantage of being simple - to be efficient and robust, it must be performed on the smallest sub-universes. Thus, with regard to live births, whenever possible a rectification must be carried out on the non-declared age of the mother, not on the basis of the live births classified according to the age of the mother but on the basis of the live births classified according to the legal nature of the birth. In a number of countries, the use of the non-declared age is especially common among *unwed* mothers.

Sequence of operations

$$E_n''^x = E_n^x * \frac{\sum_{x=ax}^{w} E_n^x}{\left(\sum_{x=ax}^{w} E_n^x\right) - ND_n}$$

2- Conversion of age pyramids drawn up on a date other than 1 January into age pyramids on 1 January

Certain European countries (United Kingdom, Ireland) and non-European countries (United States, Canada, Japan, etc.) draw up each year the estimated distribution of the resident population by sex and year of age, but the reference date to which each pyramid relates is not 1 January but another date, which is identical from one year to another: 1 July for the United Kingdom and 15 April for Ireland.

So as to process the data of all the countries using the *same* software programs, we must *first* convert the age pyramids of the United Kingdom and Ireland into pyramids on 1 January. There are two possibilities depending on whether or not the series of births by month of occurrence is available.

a) Monthly births available

The conversion made by the system is based on the *monthly* series of live births. Let us consider a year *n*: the number of residents on date *a* of year *n* (*a* being measured as a fraction of year from 1 January *n*) of persons age completed *i* is expressed as $P_i^{n,a}$. The cohort *at birth* of the members of this generation (born between date *a* of year *n*-*i*-1 and date *a* of year *n*-*i*) is estimated at $V = \int_{x=a}^{x=w} x^2 \frac{q(x)\, e^{-\int_0^x (\mu+\lambda)dx}}{1-S(w)}\, dx - m^2$ on the basis of the monthly series of live births of the period. The apparent survival ratio of this generation between the birth and the

age completed *i* is $P_i^{n,a} / N_{n-i-1,a \dashrightarrow n-i,a}$. The cohort on 1 January *n* of generation *n*-*i*-1 is expressed as P_i^n while at birth it was N_{n-i-1}. The apparent survival ratio, a_s, can be estimated by linear interpolation between two apparent survival ratios observed[13], which gives the desired cohort P_i^n when we know N_{n-i-1}.

For example, for Ireland, the age pyramid on 1 January 1990 is calculated according to this method based on the pyramids on 15 April 1989 and on 15 April 1990 (linear interpolation). It is also possible to estimate *provisionally* the pyramid on 1 January 1991 by linear extrapolation based on the previous pyramids, whereby this estimate must be revised when we know the pyramid on 15 April 1991 (which will provide a means of carrying out a *linear inter*polation and not *extrapolation*).

Sequence of operations

1- Number of residents on date *a* of year *n* (*a* being measured as a fraction of year from 1 January *n*): $P_i^{n,a}$

2- Cohort *at birth* of the members of this generation: $V = \int_{s-\infty}^{s-m} x^2 \frac{q(x)\, e^{-\int_x^s \phi(x)\,dx}}{1-S(w)}\, dx - m^2$

3- Apparent survival ratio: $P_i^{n,a} / N_{n-i-1,a \dashrightarrow n-i,a}$

4- Cohort on 1 January *n* of generation *n*-*i*-1: P_i^n (at birth it was N_{n-i-1})

5- Apparent survival ratio, a_s, estimated by linear interpolation between two apparent survival ratios observed.

6- Yields the desired cohort P_i^n when we know N_{n-i-1}

b) Monthly births not available

If the monthly distribution of births is not available for certain generations, it is nonetheless possible to carry out linear interpolations and extrapolations on the basis, for want of anything better, of a *uniform* seasonal movement of births. However, it is obviously much better to refer to the distribution of monthly births.

3- Intercensal estimates of the population

The audit of the resident population between two dates, whether it be two consecutive 1 Januarys or the dates of two successive censuses, translates the calculation equality:

Variation in the cohort of the population = Births – Deaths + Net migration

It is not uncommon for the officially approved data published by the national office to be incoherent or not to tally with the total numbers of events with all corresponding ages.

In the first case, this is generally resolved by applying the notion of *statistical adjustment*, a sort of stopgap which recognizes an irreducible *break* reckoned to occur on 31 December at midnight:

Population on 1 January *n*+1 = Population on 31 December *n* + Statistical adjustment

This solution, for all its practicality, requires an adaptation of the definition of the denominators of the occurrence rates, whether crude or by age: for the average population subject to the risk during year *n*, we use the arithmetic mean of the populations on 1

[13] The ratio $N_{n-i-1} / N_{n-i-1,a \dashrightarrow n-i,a}$ refers to the date *n*-*i*-1+*a*, the ratio $P_{i-1}^{n,a} / N_{n-i,a \dashrightarrow n-i+1,a}$ to the date *n*-*i*+*a* and the ratio P_{i-1}^{n} / N_{n-i} to the date *n*-*i*.

January *n* and on 31 December *n* and not between two successive 1 Januarys (as is the case if there is no statistical adjustment).

In the second case, there is the delicate problem of non-coherence between the official global data and the totals of the distributions of numbers of events according to age. This kind of situation arises when, following a new census, the intercensal revision of cohorts of population carried out by the national office is limited to the *total cohort* of the population and leaves the distributions by age unchanged.

Before analyzing this particular case, let us examine the difficulties with the total absence of revision of estimates of the population.

All European countries produce annual estimates of the population, but the number of countries which, once the next census is carried out, *revise* the annual estimates previously produced is rather low. This does not generally pose any major problems; however, when migratory movements are relatively high and unknown due to the fact that they are not recorded, the discrepancies between the resident population *actually* present and the estimated population can be significant[14]. A failsafe way of detecting the effects of this non-revision consists in following, over the years, the cohort of the same generation.

If the discrepancy between the estimated population and the actual population increases from one year to another, i.e. when a new census reveals a not insignificant discrepancy between the estimated population on the date of the census and the surveyed population, without an intercensal revision, the trend of all the indices involving the cohort of the population is likely to be sawtooth.

The countries that revise the cohorts of the intercensal population are yet to be identified, but the list will apparently be relatively short. Indeed, if most European countries - or at least those that still carry out a general census of the population[15] - revise the *total* cohort of their population after a new census[16], only France appears to carry out a revision of the distribution by age and sex.

Different attitudes can be adopted by Eurostat according to the policy followed by the statistics offices in this field.

a) No intercensal revision

If a country does not carry out an intercensal revision, two solutions can be adopted:

- keep only cohorts by sex and age available, which runs the risk of introducing more or less intractable discontinuities in the calculated series;
- carry out the revision of the cohorts of the population.

The revision consists, for example, in the linear interpolation over the entire intercensal period of the cohorts by age and sex between the pyramid based on the first census and the pyramid based on the second census.

[14] These differences may also be due, to a greater or lesser extent, to a lack of differential completeness of the census in relation to the previous. This is undoubtedly the reason for the difference of 480 000 persons between the population count in France in 1999 and the estimate of the population inherited from the previous census.

[15] The Netherlands and Denmark have abandoned the general census of the population.

[16] This is the case for all countries of the European Union, with the exception of Belgium and Germany.

Sequence of operations

Linear interpolation: $P_{n+x}^{g} = P_{n}^{g} * (1 + \sqrt[a]{\frac{P_{r+a}^{g}}{P_{r}^{g}}} - 1)^{x}$

b) Revision only of the total cohort of the population

When the country limits the revision of the data to the *total* cohort of the population on 1 January of each year of the intercensal period, the delicate problem arises of the non-coherence, each year, of the global official figures and of the total numbers of events by age. To overcome this difficulty, **several solutions can be adopted**, of which the relevance depends in fact on the difference between the populations estimated on the basis of the previous census and the populations revised following the new census.

- If this difference is slight and is not concentrated on the sub-populations most concerned by the phenomenon under review, this will not have any major incidence on the demographic indicators based on the old, non-revised estimate by age of the population.
- It is also possible, if this difference is not completely negligible and may result in an under-estimation or an over-estimation of the indicators, to take into account (by simple proportionality) the difference between the two successive estimates[17].

However, this solution leads to the presence in the database of two annual cohorts of population for the same date, one based on the revised official cohort and the other on the sum of the cohorts classified by age (and sex) inherited from the previous census. This may require explanatory footnotes, lead to confusions at a later date and be particularly awkward.

There are two ways to deal with this difficulty:

1. The first solution consists in retaining the revised official cohort only in the calculations of the demographic audit, whereby the total of the table distributing the resident population by age and sex is used in all other cases.
2. The second solution, *which we prefer*, consists in correcting the distribution by sex and age provided by the country in order to align its total with the annual official data.

In these cases, there are two possible procedures:

- distribute the new estimate of the cohort of the population according to the distribution by age and sex of the former estimate;
- adapt the procedure used in the event that there is no intercensal revision, i.e. interpolate annually the cohorts by age and sex between the first pyramids available after two successive censuses, then align them with the annual cohorts revised by proportional correction.

This second method is slightly more complex, but it is *without a doubt preferable*. It takes into account the ever-present possibility of discrepancies with distribution by age.

[17] A similar problem arises in countries, such as Italy, which use a definition of population to calculate the cohort of the population and its distribution which differs from that used to record events classified according to the age of occurrence.

Sequence of operations

1 - Linear interpolation: $P_{n+x}^{g} = P_{n}^{g} * (1 + \sqrt[a]{\frac{P_{r+a}^{g}}{P_{r}^{g}}} - 1)^{x}$

where P_{r}^{g} = Cohort of generation *g* in the 1st census and P_{r+a}^{g} the cohort of the same generation in the 2nd census

2 - Proportional correction: $P_{n}^{''g} = P_{n}^{g} * \frac{PR_{n}}{\sum_{g=a}^{w} P_{n}^{g}}$

where *PR* = revised total cohort of the population

4- Harmonizing the margins of the different tables concerning the same variable

We have mentioned above the difficulties that may result in different estimates of the total population. However, we may find ourselves in the same situation with the demographic events. The *total annual* number of live births, for example, can be obtained as a margin of several tables:

- distribution of births according to the month of occurrence,
- distribution of births according to sex,
- distribution of births according to the age of the mother,
- distribution of births according to their order.

With the use of computers to process civil status statistics (from the 1960s in most countries), inconsistencies between margins have generally disappeared. However, differences are recorded for the old periods.

In such a situation, in order to ensure the internal coherence of the database, one of the margins must take priority (generally the margin that will be indicated by the country as being the official number of events), which is used in the definitive demographic audit of the year under consideration. This number will then be used as the total to be reached in each of the tables dedicated to this variable, whereby the difference between this value and the margin of the table is entered and *not declared* and is subsequently processed as is, i.e. distributed according to the cohorts of each of the parts of the table considered.

We must, however, remember that this procedure leads to differences between the rectified tables and those which were initially sent by the statistics office.

C) Estimating missing data

Having to deal with missing data is a relatively common situation, whether it be a gap in a chronological series of a particular country or the absence of data for an element of a geographic universe. This can be detrimental as the non-availability of this element can prevent the calculation of a whole series of indicators. To solve this problem, it is advisable to envisage estimation methods aimed more at allowing the calculation of all the indices of which this information is only one element - unspecified but essential - of the universe rather than giving a precise estimate of the missing data.

There are many procedures that can be used, and the choice will depend above all on the type of data and how this estimate is to be used.

1- Gap in a chronological series

In the event of a gap in a chronological series, the attitude to be adopted will depend on the information available. Four cases can be identified.

a) The distribution of events by age (or by duration) is not available, but the annual total is available

In this case, which is the most favourable, there is a choice of two estimation procedures:

1. Make an estimate of the distribution of events in two phases:
   - calculate the mean of the data of the years surrounding the gap,
   - distribute according to the weight of each of the ages (or durations) the difference between the total number of events observed in the year under consideration and that obtained using the events of the surrounding years.
2. Another method that can be used, which is *a little more satisfactory* but more complex to implement, consists in using the rates of the years surrounding the gap, through a procedure that is similar to the one described above:
   - estimate the rate of the year under consideration through the average of the rates of the years surrounding the latter
   - calculate the events by age by multiplying each of the rates by age, thus estimated, by the mean population of each of these ages in the year under consideration,
   - distribute according to the weight of each of the ages (or durations) the difference between the total number of events observed in the year under consideration and that previously obtained by the sum of events by age.

This is the situation, for example, in Denmark for 1970 as regards first marriage for male and female.

Sequence of operations

1st method

1a- Linear interpolation: $E'^{g}_{n} = \frac{E^{g}_{n-1} + E^{g}_{n+1}}{2}$,

where E'^{g}_{n} = events estimated in year *n* in generation *g*

1b- Proportional correction: $E''^{g}_{n} = E'^{g}_{n} * \frac{EO_n}{\sum_{g=a}^{w} E'^{g}_{n}}$

where EO_n = *total number of events observed*

2nd method

2a- Linear interpolation: $t'^{g}_{n} = \frac{t^{g}_{n-1} + t^{g}_{n+1}}{2}$,

where t^{g}_{n} = Rate of generation *g* observed in year *n*

2b- Estimate of the events: $E'^{g}_{n} = \overline{P^{g}_{n}} * t'^{g}_{n}$

2c- Proportional correction: $E''^{x}_{n} = E'^{x}_{n} * \frac{EO_n}{\sum_{g=a}^{w} E'^{g}_{n}}$

b) The distribution of events by age (or by duration) is not available and neither is the annual total

If the total number of events is not available, one of the two above procedures will be used, with the exception of the last phase.

Sequence of operations

1st method

1a- Linear interpolation: $E'^{g}_{n} = \frac{E^{g}_{n-1} + E^{g}_{n+1}}{2}$,

where E'^{g}_{n} = events estimated in year *n* in generation *g*

2nd method

2a- Linear interpolation: $t'^{g}_{n} = \frac{t^{g}_{n-1} + t^{g}_{n+1}}{2}$,

where t^{g}_{n} = Rate of generation *g* observed in year *n*

2b- Estimate of events: $E'^{g}_{n} = \overline{P^{g}_{n}} * t'^{g}_{n}$

c) The cohort of the population by age (or by duration) is not available, but the annual total is available

To estimate the distribution by age of the population for a missing year it is advisable to:

- calculate the arithmetic mean of the cohorts belonging to the same generation on 1 January of each of the two years surrounding the gap.
- distribute the difference between the annual total observed and the annual sum of the cohorts estimated according to the annual cohorts estimated.

With this procedure for the estimation of the cohorts by generation, as is the case for the procedure described above (paragraph AI1, page 13) when only the distribution by age group is available for the year under consideration, the only difficulty lies in determining the cohorts of the first and last age: 0 and 99 and over.

The cohort at 0 years will be estimated by applying to the birth number of year *n*-1 an *apparent* survival coefficient at 0 years completed on 1 January of year *n*. The cohort of the open terminal group (99 and over) will be obtained by applying to each of the ages (98 and 99 and over) a survival probability of 1 year, then by calculating the sum of the two numbers thus obtained.

Sequence of operations

1- Linear interpolation: $P_n'^g = \frac{P_{n-1}^g + P_{n+1}^g}{2}$,

2- Proportional correction: $P_n''^g = P_n'^g * \frac{PO_n}{\sum_{g=a}^{w} P_n'^g}$

where PO_n = total observed cohort of the population

3- Cohort at 0 years: $P_n^0 = N_{n-1} * (1 - ka_0)$, where ka_0 = apparent *partial* survival coefficient

4- Cohort at 99 years and over: $P_n^{99+} = \left(P_{n-1}^{98} * (1 - q_{98})\right) + \left(P_{n-1}^{99} * (1 - q_{99})\right) + \left(P_{n-1}^{100} * (1 - q_{100})\right) + ...$

$$P_n^{99+} \approx \left(P_{n-1}^{98} * (1 - q_{98})\right) + \left(P_{n-1}^{99+} * (1 - q_{99})\right)$$

d) The cohort of the population by age (or by duration) is not available and neither is the annual total

If the total cohort is not available, the first phase of the previous procedure will be used.

Sequence of operations

1- Linear interpolation: $P_n'^g = \frac{P_{n-1}^g + P_{n+1}^g}{2}$,

2- Cohort at 0 years: $P_n^0 = N_{n-1} * (1 - ka_0)$, where ka_0 = apparent partial survival coefficient

3- Cohort at 99 years and over: $P_n^{99+} = \left(P_{n-1}^{98} * (1 - q_{98})\right) + \left(P_{n-1}^{99} * (1 - q_{99})\right) + \left(P_{n-1}^{100} * (1 - q_{100})\right) + ...$

$$P_n^{99+} \approx \left(P_{n-1}^{98} * (1 - q_{98})\right) + \left(P_{n-1}^{99+} * (1 - q_{99})\right)$$

2- Missing data at end of series (with a view to a geographic aggregation)

In the absence of provisional estimates or extrapolated estimates (within the framework of the data processing system) of the magnitude considered and if the number of geographic entities, or more precisely their weight in the universe constituted, is not too large, as far as possible an estimate should be made.

Population cohorts and demographic events are of course involved in this case.

a) The cohort of the population by age is not available

As regards the estimation of the distribution by age of a population, **various methods can be used**.

1. *The most satisfactory* method, although it also takes most time to implement, consists in using the apparent survival coefficients calculated using the last two pyramids available. This involves taking the first part of the methodology described in chapter I-A-1-b to deal with a situation in which we want to convert a pyramid by age group into an annual pyramid.
2. A faster solution consists in carrying out an extrapolation of the cohorts of each generation on the basis of the last two pyramids available, whereby the first and last ages are estimated by an extrapolation based no longer on the cohorts classified according to year of birth but on age.

3. The simplest method will be based on the simple ageing of one year of age (if the difference between the last pyramid observed and that to be estimated is only one year) of the last pyramid available. The last open group will be kept constant, and the cohort at 0 years will be estimated using the last cohort observed at this age.

If only the total cohort of the population is to be estimated, an extrapolation or a simple freeze of this cohort will be carried out.

Sequence of operations

1st method

1-. Calculate the apparent perspective survival coefficients from 1 January to the next for the last two years for which this calculation is possible:

$$(1-qa_x^{n-2}) = \frac{P_{x+1}^{n-1}}{P_x^{n-2}} \text{ and } (1-qa_x^{n-1}) = \frac{P_{x+1}^{n}}{P_x^{n-1}}$$

2-. Calculate the mean survival coefficients:

$$(1-qa_x^{n}) = \frac{(1-qa_x^{n-2})+(1-qa_x^{n-1})}{2}$$

3-. Calculate the cohorts on 1 January of year n:

$$P_x^n = P_{x-1}^{n-1} * (1-qa_{x-1}^n)$$

4-. Calculate the apparent survival coefficients of birth on the next 1 January:

$$(1-_x qa_0^{n-2}) = \frac{P_0^{n-2}}{N_{n-3}} \text{ and } (1-_x qa_0^{n-1}) = \frac{P_0^{n-1}}{N_{n-2}}$$

5-. Calculate the mean:

$$(1-qa_0^{n}) = \frac{(1-qa_0^{n-2})+(1-qa_0^{n-1})}{2}$$

6-. Calculate the cohort at 0 years on 1 January of year n:

$$P_0^n = N^{n-1} * (1-qa_0^n)$$

2nd method

1-. Extrapolation: $P_n^g = P_{n-1}^g * (1+\sqrt[2]{\frac{P_{n-1}^g}{P_{n-3}^g}} - 1)$

3rd method

1-. Reuse: $P_n^g = P_{n-1}^g$

b) The demographic events are not available

When it is the demographic events that are not available for all the countries concerned by the geographic aggregation, several procedures are possible, such as:

- carry out a linear extrapolation of the last values observed,
- extrapolate the rates of occurrence of the event considered by age, then deduce the events therefrom by applying these extrapolated rates to the population of which the cohorts by age are reckoned to be known,
- extrapolate the total rate then multiply this extrapolated indicator by the mean generation in age of the event considered, observed or itself extrapolated.

The same procedures can be easily adapted if it is the demographic indicators and not the events that are missing.

Sequence of operations

1st method

1-. Linear extrapolation of events: $E'^{x}_{n} = E^{x}_{n-1} * (1 + \sqrt[2]{\frac{E^{x}_{n-1}}{E^{x}_{n-3}}} - 1)$

2nd method

1-. Linear extrapolation of rates: $t'^{x}_{n} = t^{x}_{n-1} * (1 + \sqrt[2]{\frac{t^{x}_{n-1}}{t^{x}_{n-3}}} - 1)$

2-. Estimate of events: $E'^{x}_{n} = P^{x}_{n} * t'^{x}_{n}$

3rd method

1-. Linear extrapolation of indicators: $IC'_{n} = IC_{n-1} * (1 + \sqrt[2]{\frac{IC_{n-1}}{IC_{n-3}}} - 1)$

2-. Linear extrapolation of the mean generation: $GM'_{n} = GM_{n-1} * (1 + \sqrt[2]{\frac{GM_{n-1}}{GM_{n-3}}} - 1)$

3-. Estimate of events: $E'_{n} = GM'_{n} * IC'_{n}$

Chapter 2

Construction of tables of event occurrence

One of the main purposes of demographic analysis is to study the *occurrence of different events* within statistical universes: e.g. first marriage among single persons, the birth of a child (of a specific order or otherwise) among women or couples or death among persons of a given sex. The *intensity* of the occurrence is generally measured differentially according to age (or the period elapsed from a date taken as the origin).

Events likely to occur include repeatable events - which may occur *several* times during the life of the same individual, such as the birth of a child (of a *non-specified* order) - and *non-repeatable* events - which only occur *once* during the life of the same individual, such as first marriage, the birth of a child of a *specified biological* order or death. A non-repeatable event which *necessarily* occurs once during the life of any individual is qualified as *fatal*, such as death.

A) The different types of events

1- Repeatable events

When the event studied is *repeatable* (live births without distinction of order) or is treated as such (first marriage and divorce), the intensity of its occurrence is measured by a *rate*: ratio of the number of occurrences of the event to the number of person-years of exposure to the risk among a homogenous group of individuals *who have or have not* already experienced the event, exposed *independent* of each other and with the *same intensity* to the risk of occurrence of the event. Let us call $f(x)\mathrm{d}x$ the rate of occurrence between the ages x and $x+\mathrm{d}x$: the probability that a given individual experiences the event between the ages of x and $x+\mathrm{d}x$ is equal to $f(x)\mathrm{d}x$. The number of occurrences of the event during the life of the individual, i.e. the intensity of the phenomenon, has the mathematical expectation $I=\int_{x=\mathrm{a}}^{x=\mathrm{w}} f(x)\mathrm{d}x$, where • and • are the extreme ages when the event is likely to occur.

The age at the occurrence of the event is the continuous random variable of which the density of distribution is $\dfrac{f(x)}{\int_{x=\mathrm{a}}^{x=\mathrm{w}} f(x)\mathrm{d}x}$ and the cumulative function $F(x)=\dfrac{\int_{\mathrm{x}=\mathrm{a}}^{\mathrm{x}=x} f(\mathrm{x})\mathrm{d}\mathrm{x}}{\int_{x=\mathrm{a}}^{x=\mathrm{w}} f(x)\mathrm{d}x}$. Its mean m and its variance V are $m=\dfrac{\int_{x=\mathrm{a}}^{x=\mathrm{w}} x\,f(x)\mathrm{d}x}{\int_{x=\mathrm{a}}^{x=\mathrm{w}} f(x)\mathrm{d}x}=\mathrm{w}-\int_{x=\mathrm{a}}^{x=\mathrm{w}} F(x)\mathrm{d}x$ and $V=\dfrac{\int_{x=\mathrm{a}}^{x=\mathrm{w}} x^2\,f(x)\mathrm{d}x}{\int_{x=\mathrm{a}}^{x=\mathrm{w}} f(x)\mathrm{d}x}-m^2$ respectively. If the event is, for example, the *birth of a child* for a woman, I is, within a group made up of a large number of women, the *mean number of children per woman* and m is the *mean age at childbearing*. If the event is the *first marriage* for a single person, whereby the first marriage is treated as producing repeatable events, • is conventionally set at 50 years of age (*exact* age), I is the *proportion of ever-married persons at 50 years of age* (unit's complement of the frequency of *definitive celibacy*) and m is the *mean age at first marriage* (based on the first marriage *rates*).

2- Non-repeatable events

When the event studied is *non-repeatable* (death, marriages of single persons), the intensity of its occurrence is measured by a *probability*: ratio between the number of occurrences of the event considered and the number of person-years of exposure to the risk, among a homogenous group of individuals *who have not experienced the event* and exposed *independent* of each other and with the *same intensity* to the occurrence of the event. Let us call $q(x)\,\mathrm{d}x$ the probability of occurrence of the event between the ages of x and $x+\mathrm{d}x$: the probability that a given individual experiences the event between the ages of x and $x+\mathrm{d}x$, knowing that this event does not occur before age x, is equal to $q(x)\,\mathrm{d}x$. The probability $S(x)$ that the event occurs *after* age x, knowing that it occurred *after* age x_0, is $S(x) = S(x_0)\,\mathrm{e}^{-\int_{x_0}^{x} q(\mathrm{x})\,\mathrm{dx}}\,\mathrm{d}x$. In particular, for $x_0 = \mathrm{a}$, the minimum age at the occurrence of the event is $S(x) = \mathrm{e}^{-\int_{\mathrm{a}}^{x} q(\mathrm{x})\,\mathrm{dx}}\,\mathrm{d}x$. The probability $S(x)$ that the event occurs between the ages of x and $x+\mathrm{d}x$, knowing that it occurred *after* age x_0, is $S(x) - S(x+\mathrm{d}x) = S(x_0)\,q(x)\,\mathrm{e}^{-\int_{x_0}^{x} q(\mathrm{x})\,\mathrm{dx}}\,\mathrm{d}x$. The probability of an individual avoiding the event is $S(\mathrm{w}) = \mathrm{e}^{-\int_{\mathrm{a}}^{\mathrm{w}} q(x)\,\mathrm{d}x}$.

The age at the occurrence of the event, for those who experience it, is the continuous random variable of which the density of distribution is $\dfrac{q(x)\,\mathrm{e}^{-\int_{\mathrm{a}}^{x} q(\mathrm{x})\,\mathrm{dx}}}{1-S(\mathrm{w})}$ and the cumulative function $\dfrac{1-S(x)}{1-S(\mathrm{w})}$. Its mean m and its variance V are equal to $m = \int_{x=\mathrm{a}}^{x=\mathrm{w}} x\,\dfrac{q(x)\,\mathrm{e}^{-\int_{\mathrm{a}}^{x} q(\mathrm{x})\,\mathrm{dx}}}{1-S(\mathrm{w})}\,\mathrm{d}x = \mathrm{a} + \int_{x=\bullet}^{x=\bullet} \dfrac{S(x)}{1-S(\mathrm{w})}\,\mathrm{d}x$, $V = \int_{x=\mathrm{a}}^{x=\mathrm{w}} x^2\,\dfrac{q(x)\,\mathrm{e}^{-\int_{\mathrm{a}}^{x} q(\mathrm{x})\,\mathrm{dx}}}{1-S(\mathrm{w})}\,\mathrm{d}x - m^2$ respectively.

If the event is, for example, the *death* of an individual, $\bullet$ is the maximum duration of the human life, $S(\mathrm{w})$ is *zero* (*fatal* event) and, within a group made up of a large number of individuals, m is the *life expectancy at birth*, equal to the *mean age at death*. The *life expectancy at age x* is the mathematical expectation $e(x)$ of the number of years left to live for an individual *still living* at age x: $e(x) = \dfrac{\int_{\mathrm{x}=x}^{\mathrm{x}=\mathrm{w}} S(\mathrm{x})\,\mathrm{dx}}{S(x)} - x$.

If the event is the *first marriage* (event treated as non-repeatable) for a single person, $\bullet$ is the age set conventionally at 50 years of age to measure *definitive celibacy*, $S(x)$ is the proportion of ever-married persons at age x, $S(\mathrm{w})$ the unit's complement of the frequency of definitive celibacy, and, within a group made up of a large number of individuals, m is the *mean age at first marriage* based on the *probabilities*, calculated on the ever-married persons at 50 years of age. The mean age at first marriage calculated on the persons who marry between the ages of x and $\mathrm{w} = 50$ is $e(x) = \dfrac{\int_{\mathrm{x}=x}^{\mathrm{x}=\mathrm{w}} S(\mathrm{x})\,\mathrm{dx}}{S(x)-S(\mathrm{w})}$.

3- Observation of absolute numbers of events and of cohorts subject to the risk

The mathematical expression of the risk of occurrence of an event can be seen as measuring a *probability* of occurrence for a given individual or a *frequency* of occurrence within a group made up of a large number of *homogenous* individuals (in particular, individuals born almost simultaneously).

However, the observation data available relate to groups having a certain *amplitude*: individuals belonging to an at least *annual* age group, observed over a *duration* of at least one *year*. The previous definitions must therefore be adapted to the statistical material available.

The conventional estimates of the status of the resident population concern the cohorts of population by sex and year of age on 1 January of each year. The annual civil status statistics give the numbers of events observed during a year of observation according to the age of the person concerned[18]. This age, which is always an *integer*, can be (Figure 2) the age in years completed at the time of the occurrence of the event (numbers of events in a *square* of the Lexis diagram) or the age that the person concerned has *reached* or *will reach* during the civil year of the event (numbers of events in a *parallelogram with vertical sides* of the Lexis diagram).

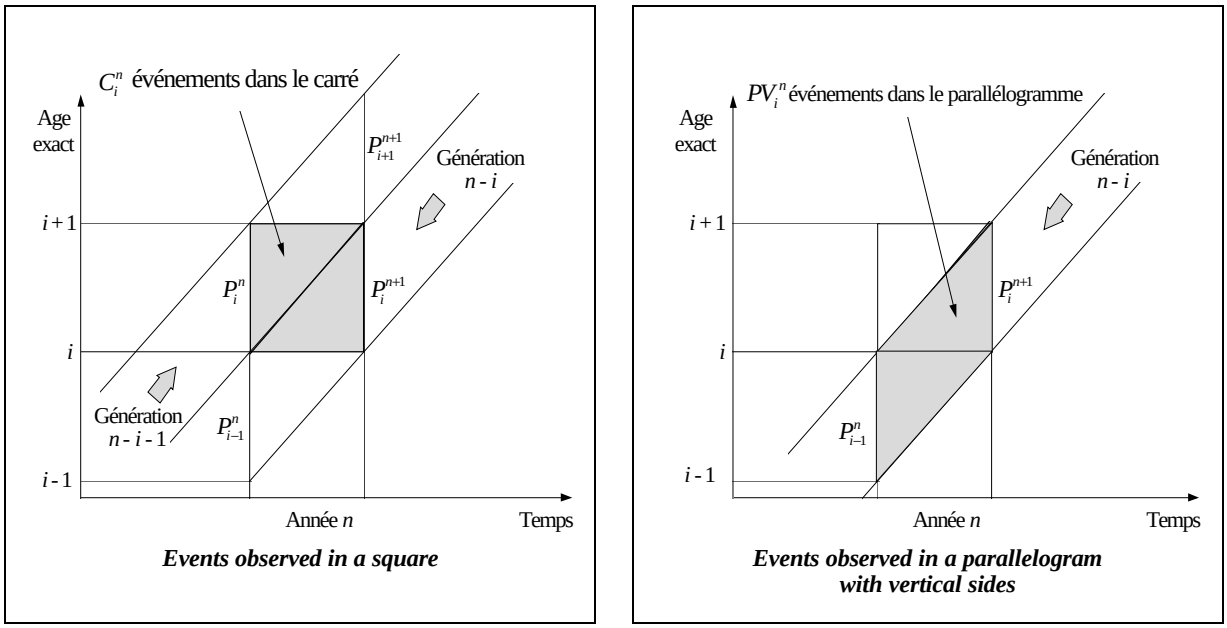


Figure 2. Absolute numbers of events and cohorts subject to the risk

[18] Or of the statistical unit concerned (for example: couple in the case of a divorce) if the event studied is undergone by statistical units other than persons.

B) Repeatable events

1- Number of person-years of exposure to the risk

This is the number of person-years of exposure to the risk which, in all rigour, constitutes the denominator of the calculation of the rates. The precise rule is to include each of the individuals who constitute the population, to which the events studied relate, for the time during which he was present during the year. It is therefore the sum of this time of presence which must be included in the denominator of rates. This is therefore a weighted mean of individuals, the weights being the fractions of year of presence.

As this calculation is becoming impossible for populations, the population observed during the year is adopted as reference population. Since the cohort of the population varies constantly during the year, the *mean population* (arithmetic mean of the cohorts of the population at the start and end of the year) is taken as the reference population. This means the population at the start and end of the year and not at the start of the next year, as to deal with calculation inconsistencies the countries may need to make statistical adjustments which create a discontinuity between 31 December and the following 1 January.

This mean population must be adapted to the definition of the age used to classify the events:

- If the events appear in the squares of the Lexis diagram (classification according to age completed), the arithmetic mean of the cohorts of the same age on two consecutive 1 Januarys is calculated;
- If the events are classified according to age reached during the year (or in difference of year (*millésime*)) in the parallelograms with vertical base of the Lexis diagram, the arithmetic mean of the cohort of the same generation on two dates is calculated.

Certain countries use as the mean population the estimate of population made on 30 June of the year, while others calculate the mean of the estimates carried out for each month (as is the case in Germany and Austria). Although the latter method is undoubtedly the most rigorous, provided, however, that all the events are observed precisely (in particular migratory movements), the difference obtained in the calculation of the demographic rates by age is so small that it is advisable to opt for the simplest method, which, in addition, provides a means of using a unique calculation method for all the countries studied.

This mean population is only the first intermediary in the rate calculation chain, and does not need to be filed since at the end of this chain the calculated rates by age will be available in the two definitions of age (age completed and age reached during the year) by correcting them taking into account the consequences of the inequality of the cohort supply as well as the random variations.

Sequence of operations

1 - Rates according to age reached during the year

arithmetic mean: $\overline{P}_i^n = \frac{P_{i-1}^n + P_i^{n+1}}{2}$,

2- Rates according to age completed

arithmetic mean: $\overline{P}_i^n = \frac{P_i^n + P_i^{n+1}}{2}$,

2- Estimate of the rates and construction of tables of occurrence of repeatable events (such as fertility)

Table 1. FINLAND, 1966. Data on fertility by age, in years completed

Age completed (years)	Pop, female on 1 January 1966	Pop, female on 1 January 1967	Births 1966	Fertility rate 1966 Raw	Fertility rate 1966 Calculated	Fertility rate 1966 Difference
14	43269	44180	7	0.0002	0.0002	0
15	45448	42987	28	0.0006	0.0006	0
16	46568	45066	220	0.0048	0.0062	-0.0014
17	48485	46231	1155	0.0244	0.0242	0.0002
18	48191	48133	2608	0.0542	0.0554	-0.0012
19	47154	47220	4306	0.0913	0.0898	0.0014
20	42481	46724	5711	0.1280	0.1196	0.0085
21	34458	41692	4919	0.1292	0.1407	-0.0115
22	32804	34402	5221	0.1554	0.1544	0.0010
23	26224	32271	4682	0.1601	0.1608	-0.0007
24	37820	25854	5092	0.1599	0.1618	-0.0018
25	26916	37519	4968	0.1542	0.1600	-0.0058
26	32002	26905	4843	0.1644	0.1554	0.0091
27	30138	31881	4526	0.1460	0.1457	0.0002
28	28446	29690	3844	0.1322	0.1315	0.0008
29	27432	28580	3258	0.1163	0.1205	-0.0042
30	27397	27273	3139	0.1148	0.1109	0.0040
31	27106	27737	2708	0.0988	0.1003	-0.0015
32	25528	26938	2312	0.0881	0.0881	0
33	27781	25619	2118	0.0793	0.0802	-0.0008
34	28684	27846	2041	0.0722	0.0720	0.0002
35	30334	28646	1841	0.0624	0.0626	-0.0002
36	30678	30412	1630	0.0534	0.0539	-0.0005
37	30264	30407	1453	0.0479	0.0473	0.0006
38	29448	30145	1239	0.0416	0.0409	0.0007
39	29976	29449	989	0.0333	0.0340	-0.0007
40	30409	29820	824	0.0274	0.0270	0.0003
41	29954	30397	633	0.0210	0.0216	-0.0006
42	30967	29790	521	0.0172	0.0166	0.0006
43	29886	30856	363	0.0120	0.0123	-0.0003
44	30702	29853	251	0.0083	0.0082	0.0001
45	31341	30561	149	0.0048	0.0048	0.0001
46	23154	31192	59	0.0022	0.0023	-0.0001
47	27659	22943	26	0.0010	0.0009	0.0001
48	27199	27633	7	0.0003	0.0003	0
49	26474	27106	6	0.0002	0.0002	0
TOTAL	1172777	1183958	77697	2.407	2.411	-0.004

MEAN AGE	30.38	30.35	27.04	27.78	27.77	-0.01

Let us take the example of fertility: birth without specification of order is a repeatable event. And let us consider the case of Finland in 1966 at 21 years completed, the year for which the numbers of events in *squares* are available. A first estimate of the fertility rate at 21 years, that we refer to as *raw*[19], is obtained by calculating the ratio of the number of births observed at 21 years completed (4 919) to the half-sum[20] of the female cohorts aged 21 years completed on 1 January (34 458) and on 31 December (41 692) of 1966 respectively, i.e. $f = \frac{4919}{(34458 + 41692)/2} = 0.1292$ (Table 1).

On the graph representing according to age the raw rates thus obtained for 1966 (Figure 3), three *couples* of anomalies can be identified (at 20-21, 25-26 and 29-30 years of age): in each couple, a first exceptional rate is immediately followed by a second, which is also exceptional, but of which the exceptional nature is the converse of the previous (here, for example, at 20 years the observed rate is too high while at 21 years it is too low).

These anomalies may be due to a *seasonal pattern* of births which is exceptional to the period in which the generations concerned are born. This is visible in the cohort at 21 years completed: 34 458 on 1 January 1966, then 41 692 on 1 January 1967 (this difference of 21% in only one year demonstrates a rapid variation in the number of births between 1944 and 1945). This can be verified in more detail by considering the temporal trend of the characteristics deduced from the statistics of monthly births (Figure 4): due to the conflict between Finland and the Soviet Union in 1940, then in 1945, the seasonal pattern of births for these two years was greatly disrupted and the raw calculation formula was inaccurate because it was based (implicitly) on the hypothesis of a *uniform* seasonal pattern of births within the couples of generations of which the rates were calculated. When such a disruption occurs, the raw method leads to biased rates at two consecutive ages, successively in one direction then in another. This is the case here with rates at 20 then 21 years (the couples of generations 1944-45 and 1945-46, including the disrupted generation of 1945) and at 25-26 years (the couples of generations 1939-40 and 1940-41, including the disrupted generation of 1940).

In contrast, this is not the case of anomalies observed at 29-30 years (the couples of generations 1935-36 and 1936-37, including the generation of 1936). However, while the anomalies linked to the disruptions of the generations of 1940 and 1945 persist in 1967 and the following years, those concerning the two couples of generations containing the generation of 1936 have disappeared (Figure 5). This is probably merely an *accidental* variation due to an observation error relating to these ages either on the female cohorts or on the number of births.

To eliminate the biases caused by exceptional seasonal patterns within certain generations, but also to *smooth* the results to limit the effect of accidental errors, the European Demographic Observatory has developed a methodology for the construction of occurrence tables. This methodology is based on the absolute numbers of events that can be observed indifferently by age completed (squares of the Lexis diagram), by age reached

[19] This is, however, the method used by most national statistics offices to calculate the tables of event occurrence.

[20] It is this estimate by half-sum of the number of person-years of exposure to the risk that creates a problem when the seasonal pattern of births within the couple of generations concerned differs significantly from uniformity, while at the same time the intensity of occurrence varies rapidly with age.

(parallelograms with vertical sides) or the two together (triangles). It gives the advanced estimates of the two types of rate according to the year of observation: by age completed

Figure 3. FINLAND, year of observation 1966
Age-specific FERTILITY rates resulting from the classical method and from ODE method
The rates computed using the classical method for couples of years containing 1936, 1940 and 1945 are shown with a circle

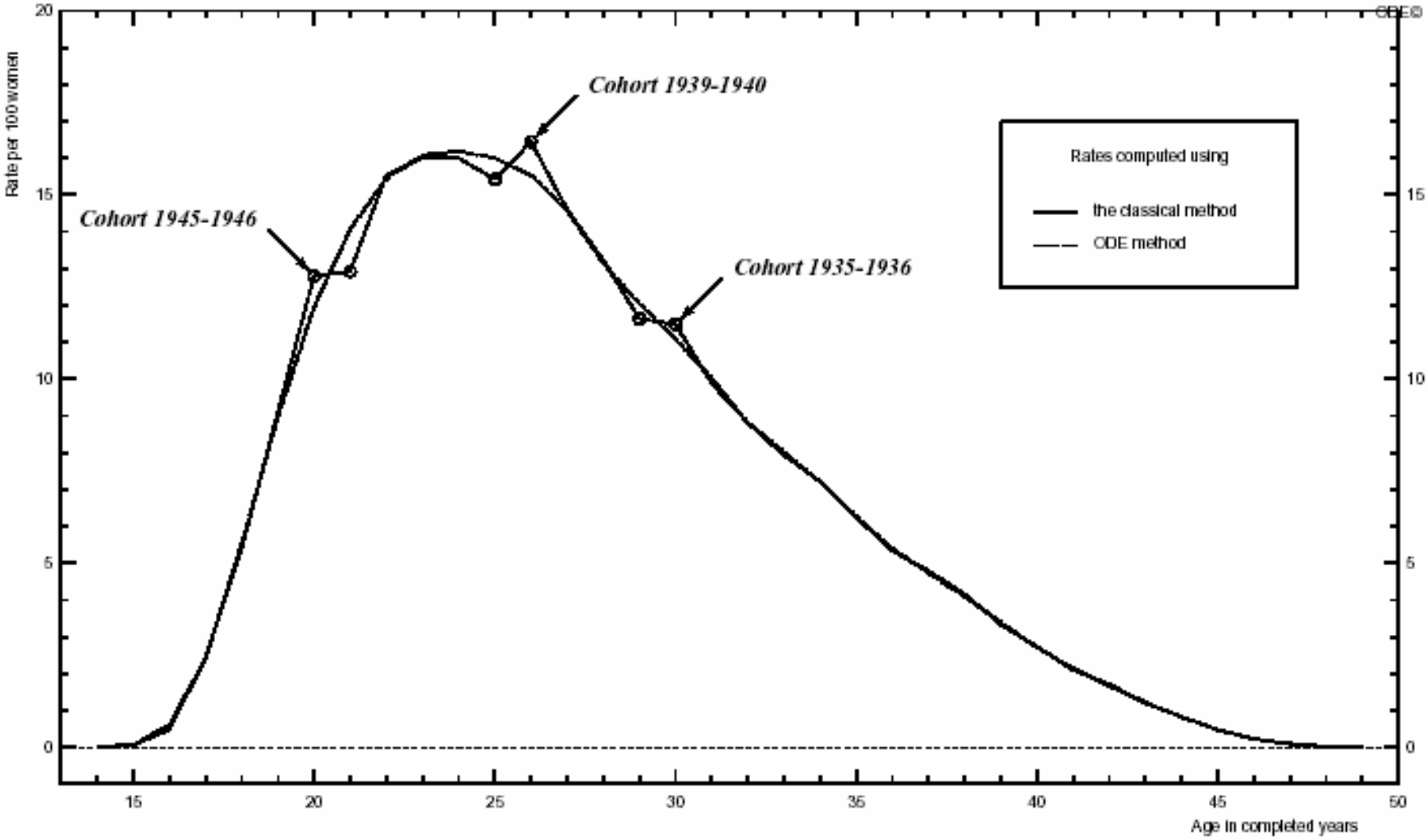


Figure 4. FINLAND. CHARACTERISTICS, at BIRTH, of BIRTH-COHORTS born from 1900 to 1998

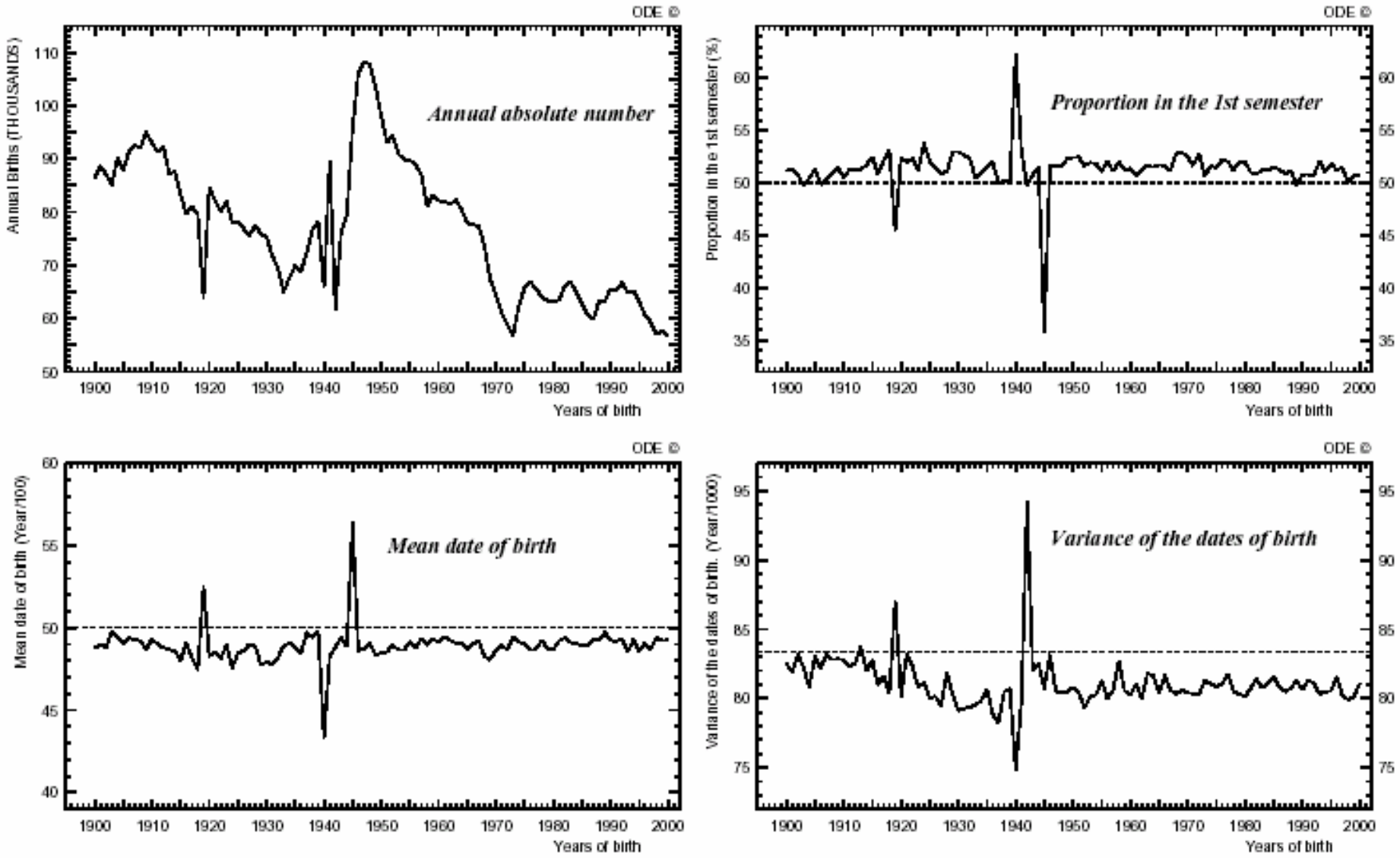
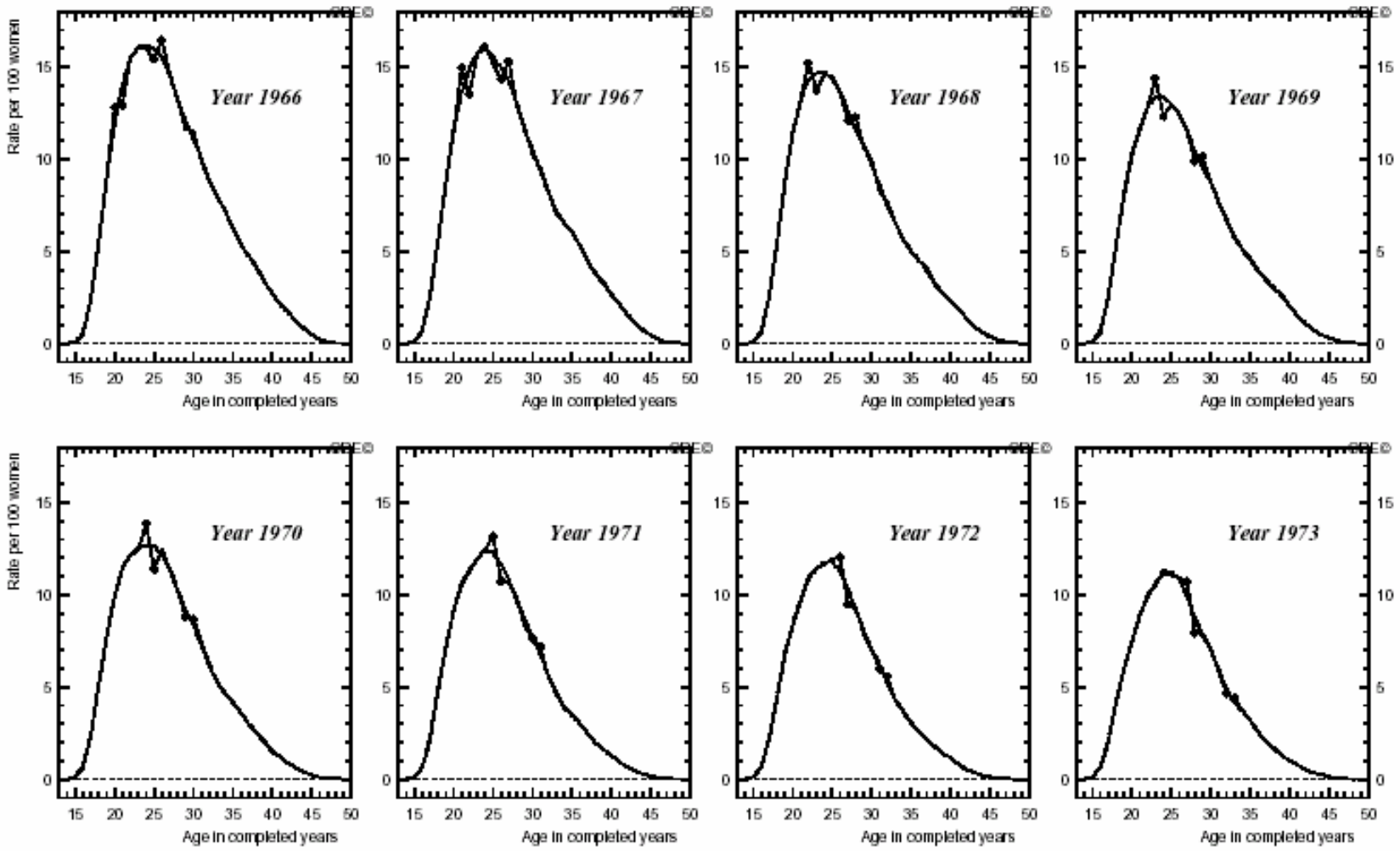


Figure 5 . FINLAND, years of observation 1966-1973
FERTILITY rates by AGE in COMPLETED YEARS computed using
the classical method (bold line) and by ODE method (dotted line)
The rates computed using the classical method for couples of years containing 1936, 1940 and 1945 are shown with a circle



(estimate of $\int_x^{x+1} f(x)\,dx$) and by age reached (estimate of $\int_{x-1/2}^{x+1/2} f(x)\,dx$). It also gives the estimate of the absolute numbers of events per triangle. Finally, it permits all the possible geographic aggregations in the Lexis diagram. Moreover, it gives the rates by age completed for the same generation, straddling two calendar years, which provides a means of calculating the cumulations at the different birthdays by generation.

The advanced estimate of the fertility rate at 21 years in 1966 is therefore 0.1407 instead of 0.1292, i.e. a *relative* difference of 8.9%. It can be demonstrated that this relative bias affecting the couples of generations 1944-45 at 21 years varies little with age, even with the phenomenon studied: it is around 9% at all ages for first marriage (Figure 6), mortality or fertility within the couple of generations 1944-45. The disruptions linked to the seasonal pattern of births therefore concern certain couples of generations and these only, but when a couple of generations is concerned, it concerns *all* ages. Furthermore, when in a given year a rate is biased for this reason, one of the two adjacent rates is affected by a bias in the opposite direction, roughly equal to an absolute value. This results in the following:

- the total rates and the mean transversal ages are affected very little by these biases (Figures 7 and 8);
- however, the longitudinal indicators are clearly regularized when they are determined using the advanced rates (Figures 9 and 10).

3- Calculation of rates (incidence rates)

This chapter concerns the general fertility rate by age, the fertility rates by age and order and the first marriage rates (male and female) by age.

For each of the definitions of age, **three methods of calculating the rates by age can be used**, *the last method being preferable* as it is the only method that provides a means of eliminating all the measurement biases.

a) Age in years completed

Raw method

The demographic rates by *age completed*, based on the absolute numbers of events observed during a given year *n* in the *squares* of the Lexis diagram, are normally obtained by calculating the ratio of the number of events E_i^n observed in the square of age completed *i* to the half-sum of the cohorts P_i^n and P_i^{n+1} of age completed *i* on 1 January and on 31 December of year *n* considered. These rates, that we shall call *raw*, are affected, as we saw in the above paragraph (B2), by *biases* due to the non-uniformity of the distribution of birthdays within each couple of consecutive generations.

Advanced method

- Constancy of the risk according to age inside the square

If the risk is *constant* inside the square, depending neither on the exact age *x* between *i* and *i*+1 nor on the time *t* during year *n*, the correct estimate of this risk is, as always, the ratio of the *absolute number of events* to the *sum of the durations of exposure to the risk*. However, even we ignore mortality and international migrations, the duration of exposure to the risk *varies* from one individual to another according to his date of birth within the couple (*n*-*i*-1, *n*-*i*): individuals born towards the start of year n-*i*-1 or at the end of year n-i are exposed to the risk during periods close to *zero*, while individuals born towards the end of year n-*i*-1 or at the start of year n-*i* are exposed to the risk for almost *one full* year.

Figure 6. FINLAND, years of observation 1964-1971
Age-specific MALE FIRST MARRIAGE rates in COMPLETED YEARS computed using the classical method (bold line) and by ODE method (dotted line)
The rates computed using the classical method for couples of years containing 1940 and 1945 are shown with a circle

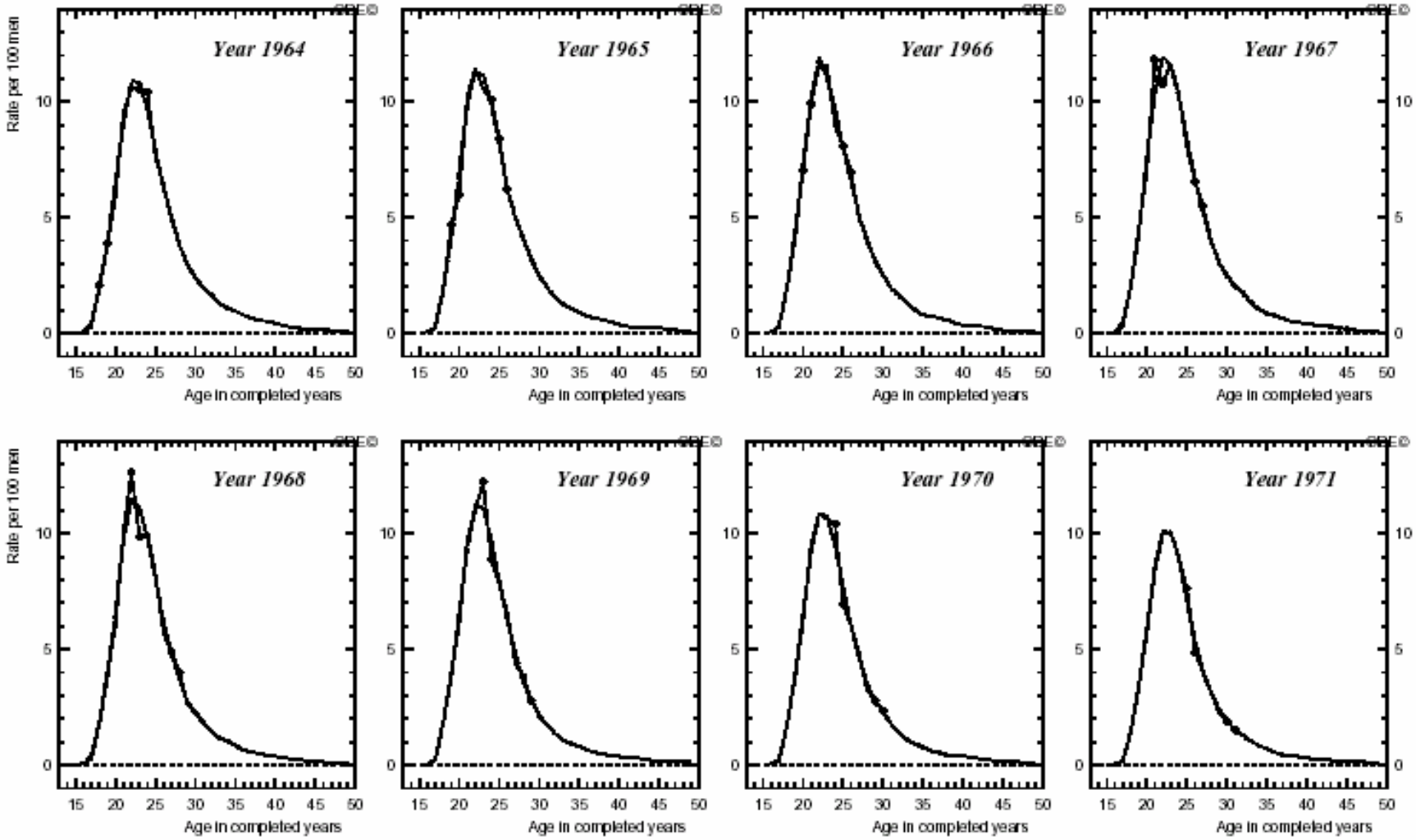


Figure 7. FINLAND, Years of observation 1941-1995. TOTAL FERTILITY RATE
Comparison between total rates derived from rates computed using the classical method and using ODE method

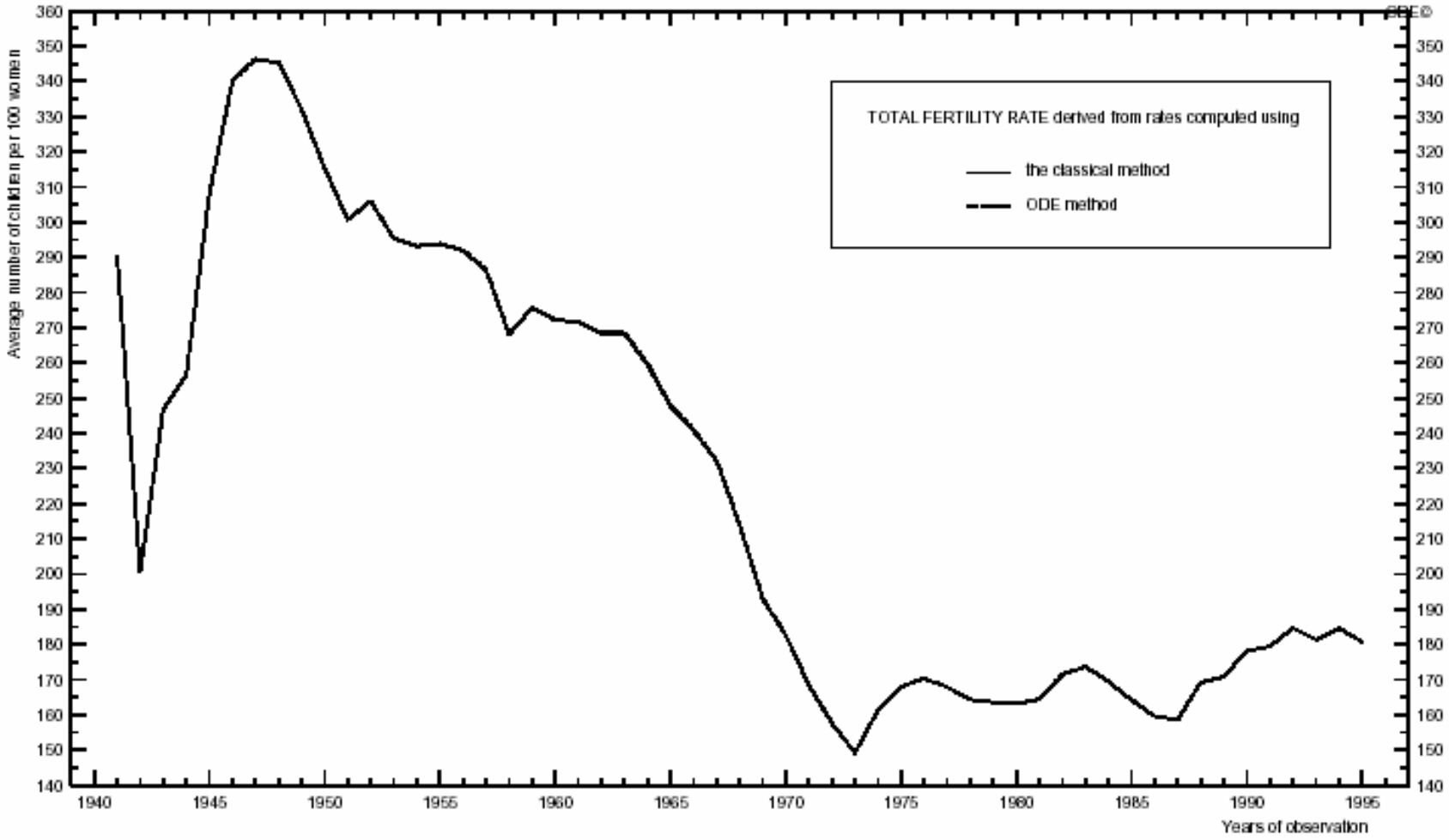


Figure 8. FINLAND, Years of observation 1941-1995. CROSS-SECTIONAL MEAN AGE at CHILDBIRTH
Comparison of mean ages at childbirth derived from rates computed using the classical method and using ODE method and based on the absolute number of events

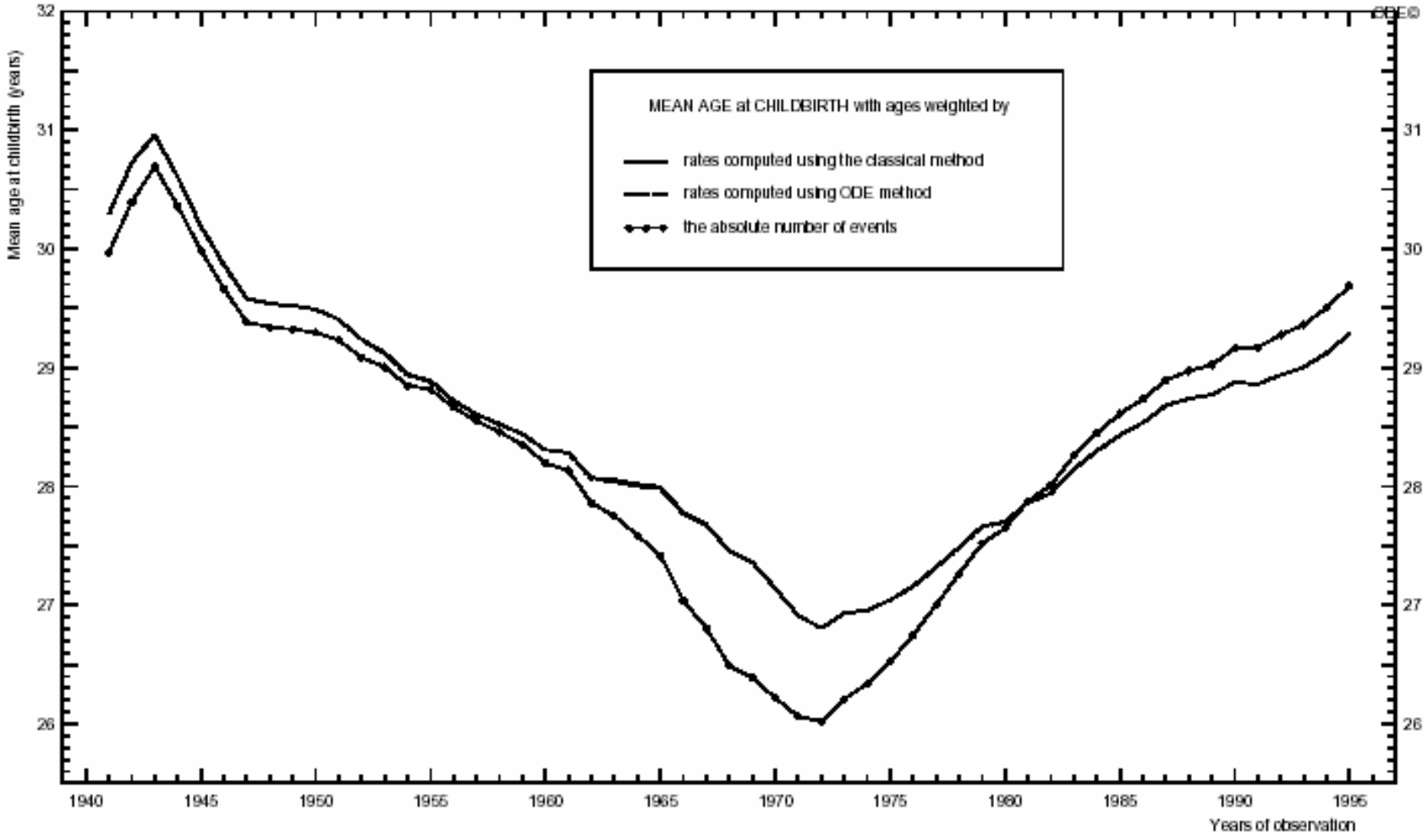


Figure 9. FINLAND, Birth cohorts 1917-1960. COHORT COMPLETED FERTILITY
Comparison of completed fertilities derived from rates computed using the classical method and using ODE method

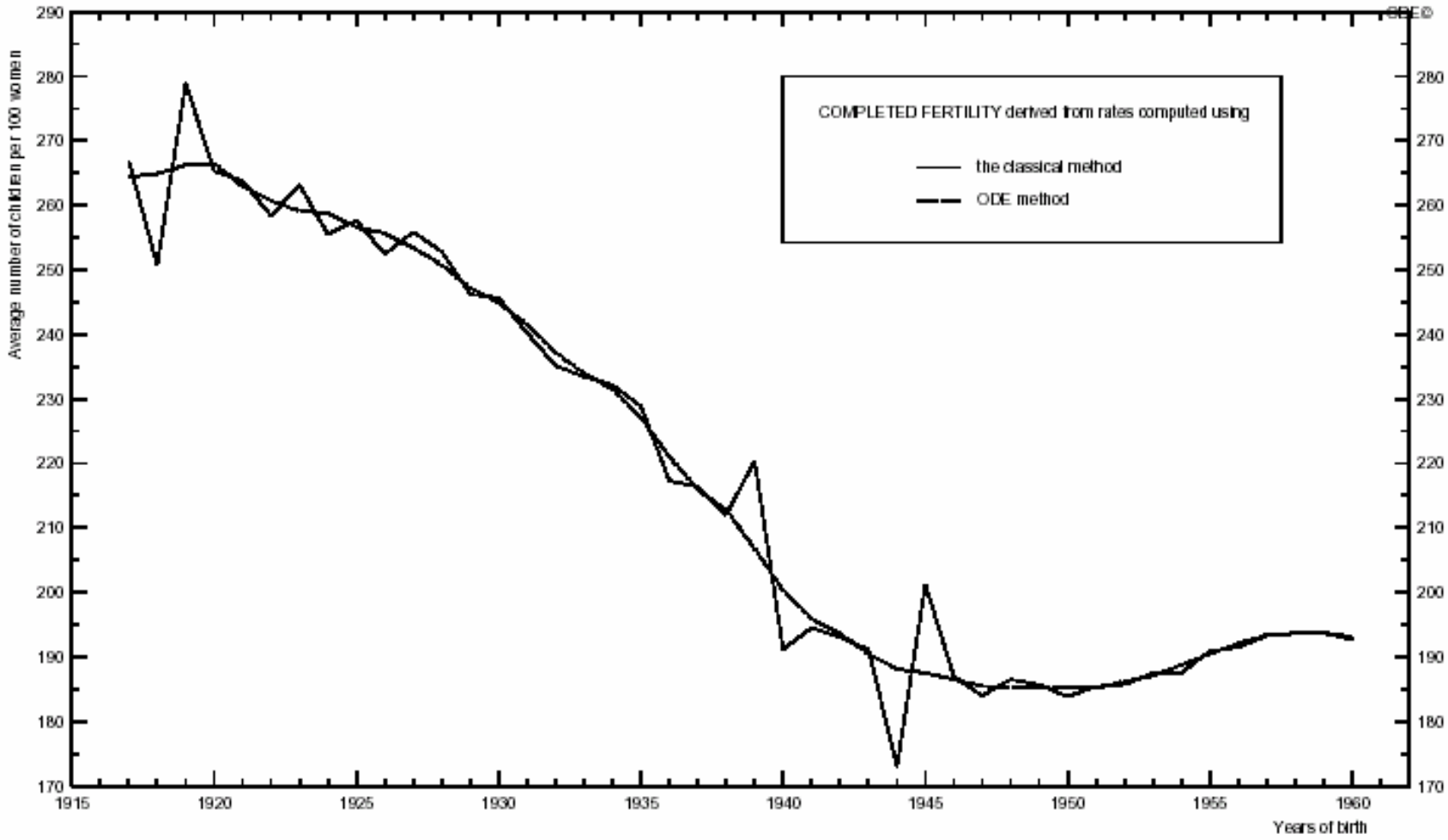
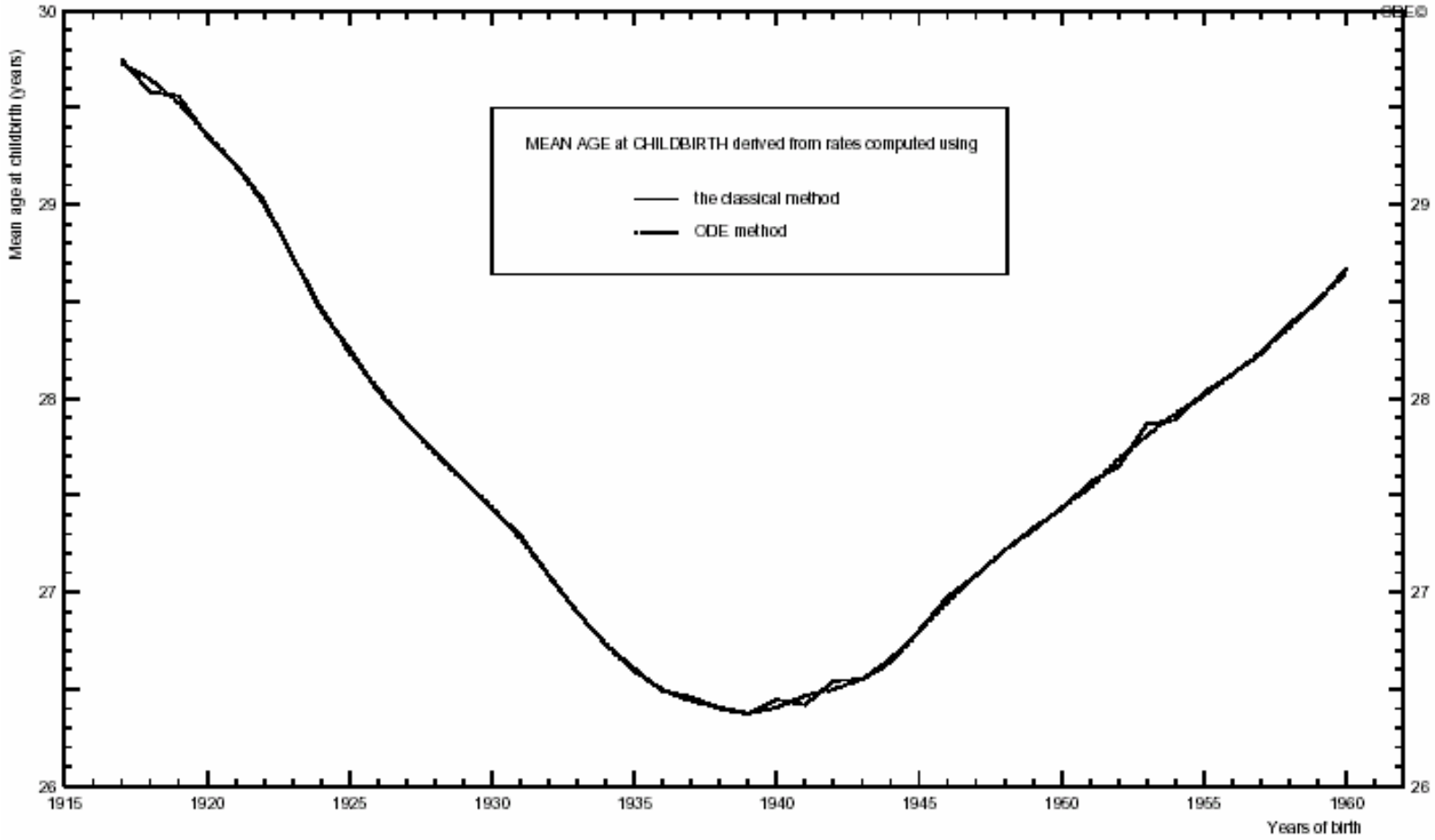


Figure 10. FINLAND, Birth cohorts 1917-1960. LONGITUDINAL MEAN AGE at CHILDBIRTH
Comparison between mean ages derived from rates computed using the classical method and using ODE method



To consider that the P_i^n individuals present on 1 January *n* have been exposed *on average* for *half a year*, as have been the P_i^{n+1} individuals present on 31 December of year *n*, we must then accept the hypothesis that the distribution of birthdays within each of the generations born in *n-i*-1 and *n-i* is *uniform*.

Based on this hypothesis, it can be demonstrated that the estimate *without bias* of the rate f_i^n at age completed *i* in year *n* is *approximately*:

$$\hat{f}_i^n = \frac{E_i^n}{P_i^n m_1 + P_i^{n+1}(1-m_2)}$$

where m_1 and m_2 are the *mean* dates of birthday *i* (counted from the start of the year and measured in years) for individuals, *present* at their birthday *i*, who were born in year *n-i*-1 and year *n-i* respectively.

Insofar as mortality and international migrations do not significantly alter the distribution of birthdays within the same generation, reference can be made to the period in which these generations were born, and the mean dates of birth m_1 and m_2 can be determined on the basis of the monthly distribution of live births during years *n-i*-1 and *n-i*. We can also refer to a recent census which gives the distribution of the resident population by year and *month* of birth (or even *day* of birth).

The bias which affects all the rates of the same couple of generations is thus a *purely statistical* bias, which is roughly *constant* in *relative* value, *independent* of both the age and the phenomenon studied, whether it concerns fertility or first marriage. With regard to mortality, the bias which affects the *probabilities* is roughly the same, in relative value, as that which affects the fertility or first marriage rates.

- Taking into account the variability of the risk according to age inside the square

If we take into account the fact that the risk $f(x)$ *varies* inside the square (*i*• *x*• *i*+1) along with mortality and international migrations, it can be demonstrated that, based on the hypotheses specified below and by calling f_i, f_i' and f_i'' the values of $f(x)$ and of its first two derivatives in $x = i + \frac{1}{2}$, the rates f_i^n to be estimated at age completed *i* is:

$$f_i^n = f_i + \frac{f_i''}{24}$$

while the absolute number of events E_i^n observed in the square[21] is equal to:

[21] The following equation is the sum of the equations corresponding to each of the triangles:

$$E_1 = N_1\left[f m_1 + \frac{f'}{2}\left(m_1 - m_1^2 - V_1\right)\right] + \frac{f''}{2}\left(\frac{m_1}{4} - \frac{m_1^2 + V_1}{2} + \frac{m_1^3}{3} + m_1 V_1 + \frac{m_1}{3}\right) - \frac{s_1}{6}\left(f + \frac{f''}{40}\right)$$

$$E_2 = N_2\left\{ f(1-m_2) - \frac{f'}{2}\left[m_2 - m_2^2 - V_2\right] + \frac{f''}{2}\left[\frac{1}{12} - \frac{m_2}{4} + \frac{m_2^2 + V_2}{2} - \frac{m_2^3}{3} - m_2 V_2 - \frac{m_2}{3}\right]\right\} + \frac{s_2}{6}\left[f + \frac{f''}{40}\right]$$

We use these formulae to estimate the number of events in each triangle and to deduce from this the number of events in the vertical-base parallelogram, enabling us to calculate occurrence rates to be calculated according to the age reached that year.

$$E_i^n = \left(P_i^n + \frac{s_1}{2}\right)\left\{f m_1 + \frac{f'}{2}\left(m_1 - m_1^2 - V_1\right) + \frac{f''}{2}\left(\frac{m_1}{4} - \frac{m_1^2 + V_1}{2} + \frac{m_1^3}{3} + m_1 V_1 + \frac{\mathfrak{m}_1}{3}\right)\right\}$$
$$+\left(P_i^{n+1} - \frac{s_2}{2}\right)\left\{f(1-m_2) - \frac{f'}{2}\left[m_2 - m_2^2 - V_2\right] + \frac{f''}{2}\left[\frac{1}{12} - \frac{m_2}{4} + \frac{m_2^2 + V_2}{2} - \frac{m_2^3}{3} - m_2 V_2 - \frac{\mathfrak{m}_2}{3}\right]\right\} \quad (1)$$
$$+\frac{s_2 - s_1}{6}\left(f + \frac{f''}{40}\right)$$

where we have omitted (to simplify the equation) the index *i* of f, f' and f'' and where s_1 and s_2 are the apparent net migration densities in the upper and lower triangles of the square respectively:

$$s_1 = P_{i+1}^{n+1} - P_i^n, \quad s_2 = P_i^{n+1} - P_{i-1}^n$$

while m_1, V_1 and $\mathfrak{m}_1$ are the mean, the variance and the central moment of order 3 respectively of the distribution of birthdays within the generation born in *n*-*i*-1, and m_2, V_2 and $\mathfrak{m}_2$ are the analogous quantities for the generation born in *n*-*i*. When within a generation the distribution of birthdays is *uniform*, i.e. when the *density* of the life lines is *constant*, the values of m, V and $\mathfrak{m}$ are equal to $\frac{1}{2}$, $\frac{1}{12}$ and 0 respectively.

These results are based on the following hypotheses:

- migrants, prior to their emigration or after their immigration, as well as persons who have died, prior to their death, are exposed to the *same* risk as individuals who are constantly present;
- in the Lexis diagram, the points representing entry (by immigration) and exit (by emigration or death) are distributed *uniformily* inside the same parallelogram with vertical sides, specific to the same generation during the year considered;
- the function of risk depends on the age but not on the *time* in the year: the phenomenon studied is assumed not to have *seasonality*;
- the function of risk $f(x)$ varies *slowly* over an age interval of several consecutive years, so that the polynomial developments of $f(x)$ at order 2 are satisfactory;
- the quantities m_1, V_1 and $\mathfrak{m}_1$ as well as m_2, V_2 and $\mathfrak{m}_2$ are *known* (evaluated on the basis of the monthly distribution of births during year n-*i*-1 and year n-i or on the basis of a recent census).

The rates f_i^n are estimated by successive *iterations*. At the first iteration, we assume f_i' and f_i'' are *zero* for any i and we deduce E_i^n, P_i^n, P_i^{n+1}, P_{i-1}^n and P_{i+1}^{n+1} using equation (1), f_i, then $f_i^n = f_i$. At each subsequent iteration, we estimate f' and f'' by a parabolic fit of the values of f obtained at the *previous* iteration over, for example, 5 consecutive points:

$$f_i' = \frac{1}{5}\left(f_{i+2} - f_{i-2}\right) + \frac{1}{10}\left(f_{i+1} - f_{i-1}\right)$$

$$f_i'' = \frac{2}{7}\left(f_{i+2} + f_{i-2} - f_i\right) - \frac{1}{7}\left(f_{i+1} + f_{i-1}\right)$$

and we deduce E_i^n, P_i^n, P_i^{n+1}, P_{i-1}^n and P_{i+1}^{n+1}, again using equation (1), f_i, then $f_i^n = f_i + \frac{f_i''}{24}$. The iterations are stopped when the estimates f_i^n are stabilized for any i. The advantage of this procedure is that the data can be *smoothed*, which attenuates the effects of accidental variations.

If we ignore the quantities s_1 and s_2 as well as the derivatives of order 1 and 2 of $f(x)$, one single iteration is required, and the estimation of the rates f_i^n is that indicated above:

$$f = \frac{E_i^n}{P_i^n m_1 + P_i^{n+1}(1 - m_2)}$$

while the conventional estimation is:

$$f = \frac{E_i^n}{(P_i^n + P_i^{n+1})/2}$$

As regards the bias affecting the conventional estimation:

- when P_i^n and P_i^{n+1} are adjacent, m_1 and m_2 are different (extremely uncommon situation, as, when m_1 and m_2 are different, P_i^n and P_i^{n+1} are generally also different);
- when P_i^n and P_i^{n+1} are different, m_1 and m_2 differ by ½.

When the distribution of births during the couple of years is influenced virtually only by the *seasonal* variations of natality, experience shows that the bias is *negligible*. However, for the couples marked by *sudden* variations in natality, for example at the end and, more especially, at the *start* of a war, the bias can reach significant values and lead to major errors in relation to that of the variation, from one age to another, in the rates observed in the same year and, more especially, in relation to the variation, from one year to another, in the rates at the same age. The bias therefore affects the couple at *all* ages, i.e. throughout its existence, remaining roughly at the same *relative* valu e. It follows that the estimate of the *longitudinal* characteristics of the couple of generations is affected by the relative bias common to the rates at each of the ages.

It is therefore desirable that the software applications be developed on the basis of the methodology described above and estimate, in a unique manner for all the countries, the rates at all ages and for all years. They should also be able to calculate, optionally, the rates according to the conventional method (raw rates). When the characteristics m, V and $\mathtt{m}$ of a couple of generations are not available, we must use, *for want of something better*, values equal to $\frac{1}{2}$, $\frac{1}{12}$ and 0 respectively.

b) Age reached during the year

Raw method

The demographic rates by age *reached during the year*, based on the absolute numbers of events observed during a year *n* given in the *parallelograms with vertical base* of the Lexis diagram, are normally obtained by calculating the ratio of the number of events E_i^n observed in the parallelogram of age reached *i* to the half-sum of the cohorts P_{i-1}^n and P_i^{n+1} of age completed *i* on 1 January and *i+1* on 31 December of year n considered. These *raw* rates are also affected by *bias*, but this is considerably smaller than that of the rates by age completed.

Advanced method

When the events are classified in a parallelogram with vertical sides, the formulas are much simpler as only one single generation is involved.

The rate f_i^n to be estimated at the age reached *i* is always:

$$f_i^n = f_i + \frac{f_i''}{24}$$

while the absolute number of events E_i^n observed in the parallelogram[22] becomes equal to:

$$E_i^n = \left(P_{i-1}^n + \frac{s}{2}\right)\left\{f + f'\left(m - \frac{1}{2}\right) + \frac{f''}{2}\left(\frac{1}{3} - m + m^2 + V\right)\right\} + s\frac{f'}{6} \quad (2)$$

where s is the apparent net migration density in the parallelogram of age reached *i*:

$$s = P_i^{n+1} - P_{i-1}^n$$

while m and V are the mean and the variance respectively of the distribution of birthdays within the generation born in *n-i*.

The estimation of the rates f_i^n is always carried out by successive *iterations*.

4 Conversions of events and rates in another Lexis diagram

The procedure used consists of estimating numbers of events in each of the triangles making up the original diagram, starting with the occurrence function calculated in this last diagram, applied in each of the specific equations of the various triangles.

The following phase therefore consists of aggregating the appropriate triangles in the Lexis diagram and then calculating the rates in these new diagrams with the help of the specific equation.

Sequence of operations

1 – Raw rates

- calculation of the mean population by arithmetic mean:

$\overline{P}_i^n = \frac{P_i^n + P_i^{n+1}}{2}$ or $\overline{P}_{n-i}^n = \frac{P_{i-1}^n + P_i^{n+1}}{2}$, according to the case

- calculation of the rate:

$f_i^n = \frac{E_i^n}{\left(P_i^n + P_i^{n+1}\right)/2}$ or $f_{n-g}^n = \frac{E_i^n}{\left(P_{i-1}^n + P_i^{n+1}\right)/2}$, according to the case

2- Advanced rates (age completed)

[22] Based on the addition of the two triangles:

$$E_1 = N\left[f\,m - \frac{f'}{2}\left(m^2 + V\right) + \frac{f''}{6}\left(m^3 + 3mV + \mathbb{m}\right)\right] - s\left(\frac{f}{6} - \frac{f'}{12} + \frac{f''}{40}\right)$$

$$E_2 = N\left\{f(1-m) + \frac{f'}{2}\left[(1-m)^2 + V\right] + \frac{f''}{6}\left[(1-m)^3 + 3(1-m)V - \mathbb{m}\right]\right\} + s\left(\frac{f}{6} + \frac{f'}{12} + \frac{f''}{40}\right)$$

with $N = \left(P_{i-1}^n + \frac{s}{2}\right)$

a) constancy of the risk

- calculation of the characteristics at birth: m_1 *and* m_2,
- calculation of the rate:

$$\hat{f}_i^n = \frac{E_i^n}{P_i^n m_1 + P_i^{n+1}(1-m_2)}$$

b) variable risk

- calculation of the characteristics at birth: m, V and $\mathtt{m}$, in their absence use values equal to $\frac{1}{2}$, $\frac{1}{12}$ and 0 respectively.
- estimation of the rate, $f_i^n = f_i + \frac{f_i''}{24}$, calculated by successive *iterations*

- at the 1st iteration: - we assume f_i' and f_i'' (first and second derivatives) are *zero* for any i

- we deduce from E_i^n, P_i^n, P_i^{n+1}, P_{i-1}^n and P_{i+1}^{n+1}, using the equation (1), f_i, then $f_i^n = f_i$

- at each subsequent iteration: - we estimate f' and f'' by a parabolic fit of the values of f obtained in the *previous* iteration

- we deduce from E_i^n, P_i^n, P_i^{n+1}, P_{i-1}^n and P_{i+1}^{n+1}, again using the equation (1), f_i, then $f_i^n = f_i + \frac{f_i''}{24}$

- Estimate of the events in the triangles making up the squares in order to deduce these in the parallelograms with vertical base.

3- Advanced rates (age reached)

- calculation of characteristics at birth: m and V (in their absence use $\frac{1}{2}$ and $\frac{1}{12}$).
- estimation of the rate, $f_i^n = f_i + \frac{f_i''}{24}$, calculated by successive *iterations*

- at the 1st iteration: - we assume f_i' and f_i'' (first and second derivatives) are *zero* for any *i*

- we deduce from E_i^n, P_i^{n+1} and P_{i-1}^n, using the equation (2), f_i, then $f_i^n = f_i$

- at each subsequent iteration: - we estimate f' and f'' by a parabolic fit of the values of f obtained at the *previous* iteration

- we deduce from E_i^n, P_i^{n+1} and P_{i-1}^n, again using the equation (2), f_i, then $f_i^n = f_i + \frac{f_i''}{24}$

- Estimate of the events in the triangles making up the parallelograms in order to deduce these in the squares.

C) Non-repeatable events

1- Estimating probabilities and constructing tables of occurrence of non-repeatable events (such as mortality)

The statistical data available to analyze the occurrence of non-repeatable events are presented in the same way as for repeatable events: absolute numbers of events in parallelograms with vertical sides, in squares or the two simultaneously, i.e. in triangles. However, the construction method for tables of occurrence is different. In the case of repeatable events, the aggregation of the rates is carried out by addition, while for non-repeatable events the one's complements of the occurrence probabilities are sequenced by multiplication.

As is the case for the calculation of rates, two cases can be identified, according to the definition of the age used to classify the events.

a) Age in completed years

Raw method

With regard to deaths, most European countries have had data in *triangles* for many years (for example, France since 1907).

Let us call P_i^n the cohort on 1 January n of the population of a given sex at age completed i and C_i^n the number of deaths in the corresponding square at age completed i in year n.

The *raw* method, which is applied to the data in triangles or in squares, consists in estimating the probability of dying ${}_1Q_i^n$ [23] at age completed i as follows: we first estimate the mortality *rate* t_i^n as if it were a repeatable event by $t_i^n = \frac{C_i^n}{\left(P_i^n + P_i^{n+1}\right)/2}$, then the probability itself by ${}_1Q_i^n = 1 - \mathrm{e}^{-t_i^n} \approx \frac{t_i^n}{1 + t_i^n/2}$ (the close expression is valid only when the rate t_i^n is small). The raw method is based implicitly on two hypotheses: hypothesis of *uniformity* of the seasonal pattern of births during the couple of years n-i-1 and n-i as in the case of repeatable events and hypothesis of *slight* variations in the instantaneous probability of dying between the ages of i and i+1.

For this reason, a more advanced methodology must be used which does not assume that these hypotheses are satisfactory and which in addition carries out *smoothing*, which is even more desirable with regard to mortality than with regard to fertility, as the "risks" - and therefore the absolute numbers of events - are often much lower. In particular, as the observed data are often of mediocre quality at old ages and that, in parallel, the cohorts are reduced whereby the probabilities become erratic, the methodology of the EDO uses, as an option, the additional hypothesis that beyond a certain age (for example, 85 years) the *logarithm* of the probability of dying varies *linearly* with the age: we therefore adjust a straight line, for example to the *ten* previous data (here: from 70 to 89 years of age) and we read the next data (here: from 90 years of age) on the adjusted straight line[24]. This

[23] The lower left index 1 expresses the fact that it is a probability over an interval of *one* year: between the exact ages i and i+1.

[24] The calculation of life expectancy at all ages takes account of the fact that at 99 years the remaining life expectancy is not zero or *a priori* fixed: it is equal to the remaining life expectancy based on the indefinitely extrapolated probabilities (linear extrapolation of the probability logarithms).

methodology must yield, for each age *i* from 0[25] to 99 years, the probability of dying ${}_1Q_i^n$ at years completed *i*, the number of survivors S_i^n at the exact age *i* and the life expectancy remaining at the exact age *i*.0

Advanced method

This methodology used for the calculation of probabilities is exactly the same as the methodology described above for the calculation of incidence rates. The procedure is identical; only the formulas change.

In the case of events classified according to age in completed years (squares of the Lexis diagram) we can, as for the calculation of rates, choose between two hypotheses concerning the trend of the risk inside the square: constant or variable risk.

We will not go into the details of the calculation as this would involve repeating what has been said in the previous paragraph; we will confine ourselves to giving the formulas to be used and to indicating a few particularities[26].

- Constancy of the risk according to age inside the square

After calculating the characteristics at birth of the generations concerned, the rate of the next formula must be calculated, which shows that the approximate effect of migrations on the number of deaths of a square in the Lexis diagram is linked to the difference in the net migrations of the two triangles making up this square:

$$\hat{q}_i^n = \frac{E_i^n}{P_i^n m_1 + P_i^{n+1}(1-m_2) + \frac{s_2 - s_1}{6}}$$

and this rate must be converted into a probability:

$${}_1Q_i^n = 1 - \mathrm{e}^{-\hat{q}_i^n} \approx \frac{\hat{q}_i^n}{1 + \hat{q}_i^n / 2}$$

- Taking into account the variability of the risk according to age inside the square

In this case, the number of events in the square takes the following form:

$$SQ_i^n = \left(P_i^n + \frac{s_1}{2}\right)\left[\mathrm{e}^{qm_1 + \frac{q'}{2}[m_1(1-m_1)-V_1] + \frac{V_1}{2}q^2} - 1\right] + \left(P_i^{n+1} - \frac{s_2}{2}\right)\left[1 - \mathrm{e}^{-q(1-m_2) + \frac{q'}{2}[m_2(1-m_2)-V_2] + \frac{q^2}{2}V_2}\right] + \frac{s_1}{2}\left[1 - \mathrm{e}^{\frac{q}{3} + \frac{q^2}{36}}\right] + \frac{s_2}{2}\left[1 - \mathrm{e}^{-\frac{q}{3} + \frac{q^2}{36}}\right] \quad (3)$$

And to estimate the instantaneous probability *q*, or mortality rate, we use the same iterative procedure as is used to estimate the incidence rate of repeatable events. However, there are two possibilities:

- either we consider *q* to be constant:
  - at the first iteration $q_i' = 0$,
  - at subsequent iterations $q_i' = \frac{q_{i+1} - q_{i-1}}{2}$

[25] At ages 0 and 1 year, the methodology differs from that applied to other ages due to the *rate* of variations of the instantaneous probability of dying with age. In addition, it should be mentioned that the probability of dying at 0 years is usually referred to as the *infant mortality rate*.

[26] A precise description of this method, used by the EDO, is given in the article by Gérard Calot and Ana Franco in "The construction of life tables" which appears in the work published by Guillaume Wunsch, Michel Mouchart and Josianne Duchêne (editors), *Life Tables: Data, Methods, Models*, Kluwer, 2001, pp. 31-75, which is reproduced in the annex to this report.

- or it is the logarithmic derivative $\frac{q_i'}{q_i}$ which is constant:

- at the first iteration the derivative is equal to 0,
- at subsequent iterations to $\frac{Log(q_{i+1}) - Log(q_{i-1})}{2} = 0{,}5 * \left(\frac{Log(q_{i+1})}{Log(q_{i-1})} \right)$

We then calculate the finished probability by:

$$ {}_1Q_i^n = 1 - e^{-q} \approx \frac{q}{1 + q/2} $$

If the number of deaths is known only inside the square and not in each of the triangles, the above iterative procedure must be preceded by a step, which is also iterative, during which the number of deaths in each of the triangles will be estimated based on the following formulas:

Lower triangle:

$$ \begin{aligned} D_2 &= N_2 - P_2 + \frac{S_2}{2} \\ &\approx N_2 \left\{ 1 - e^{-q(1-m_2) + \frac{q'}{2} m_2 (1-m_2)} \left[1 + \frac{V_2}{2} (q^2 - q') \right] \right\} + \frac{S_2}{2} \left[1 - e^{-\frac{q}{3}} \left(1 + \frac{q^2}{36} \right) \right] \end{aligned} $$

Upper triangle:

$$ \begin{aligned} D_1 &= P_1 - N_1 + \frac{S_1}{2} \\ &\approx N_1 \left\{ e^{q m_1 + \frac{q'}{2} m_1 (1-m_1)} \left[1 + \frac{V_1}{2} (q^2 - q') \right] - 1 \right\} - \frac{S_1}{2} \left[e^{\frac{q}{3}} \left(1 + \frac{q^2}{36} \right) - 1 \right] \end{aligned} $$

b) Age reached during the year

Raw method

In the case of the parallelogram with vertical sides, the formula to calculate the rate is similar to that used to calculate deaths classified according to age in completed years:

$$ t_i^n = \frac{C_i^n}{\left(P_i^n + P_{i+1}^{n+1} \right) / 2} $$

And the rate is converted into a probability in the same way, the only difference being that the probability is defined between the exact ages of *i-0.5* and *i+0.5:*

$$ {}_1Q_{i-0,5}^n = 1 - e^{-t_i^n} \approx \frac{t_i^n}{1 + t_i^n / 2} $$

Advanced method

Once again the procedure is the same; only the equation giving the events in the parallelogram with vertical[27] sides differs:

---

[27] In this case the equations corresponding to each triangle are:

$$ E_1 = N \left[e^{q m - \frac{q'}{2} \left(m^2 + V \right) + \frac{q^2}{2} V} - 1 \right] + \frac{s}{2} \left[1 - e^{\frac{q}{3} - \frac{q'}{6} + \frac{q^2}{36}} \right] $$

$$ E_2 = N \left[1 - e^{-q(1-m) - \frac{q'}{2} \left[(1-m)^2 + V \right] + \frac{q^2}{2} V} \right] + \frac{s}{2} \left[1 - e^{-\frac{q}{3} - \frac{q'}{6} + \frac{q^2}{36}} \right] $$

with $N = \left[\frac{P_{i-1}^n + P_i^{n+1}}{2} + \frac{D_{1,i}^n + D_{2,i}^n}{2} \right]$

$$VSP_i^n = \left[\frac{P_{i-1}^n + P_i^{n+1}}{2} + \frac{D_{1,i}^n + D_{2,i}^n}{2}\right]\left[1 - e^{-q - \frac{q'}{2}(1-2m)}\right] e^{qm - \frac{q'}{2}(m^2+V) + \frac{q^2}{2}V} + \frac{s}{2}\left[2 - e^{\frac{q}{3} - \frac{q'}{6} + \frac{q^2}{36}} - e^{-\frac{q}{3} - \frac{q'}{6} + \frac{q^2}{36}}\right] \quad (4)$$

We estimate the instantaneous probability or rate using the above equation, by opting to consider as constant the instantaneous probability or its logarithmic derivative. The rate is then converted into a probability.

The disadvantage of these (*perspective*) probabilities is that they are not calculated between birthdays but between two successive first Januarys; they are therefore shifted by one half-year of age. For most ages, except the very young, perspective probabilities can be converted into probabilities between birthdays by linear interpolation or, preferably, by logarithmic interpolation. However, it is preferable to use the procedure described above to estimate the deaths in each of the triangles and to deduce from this a table between birthdays.

c) Specificity of mortality at 0 and 1 year

Everything that has been described in this chapter devoted to the estimation of probabilities applies whatever the phenomenon studied (fertility, nuptiality or mortality) and the age considered, with the exception of mortality at 0 years and, to a lesser extent, 1 year. The very rapid decline in the instantaneous probability, and thus of the function of risk, from birth thwarts the polynomial approximation used.

We can therefore estimate the probability of dying at age 0, based on the two hypotheses of zero migrations in the square and of identical distributions of birthdays of the generations born in *n-1* and *n*, by:

$$\hat{Q} = 1 - \left(1 - \frac{D_1}{P_1}\right)\left(1 - \frac{D_2}{N_2}\right)$$

where D_1 and D_2 = deaths of upper and lower triangles,

P_1 = population of 0 years completed on 1 January of the year considered,

and N_2 = births of the year considered.

In the case of migrations in this first square, the above formula can be rewritten as follows:

$$\hat{Q} = 1 - (1-a)(1-b)$$

To estimate *a* (which relates to the upper triangle) and *b* (the lower triangle), two hypotheses can be used as regards the place of these migrations:

- for the upper triangle, the first hypothesis assumes that the migrations of year *n* of the generation born in *n-1* take place on 1 January of year *n* (a_1); the second, that the migrations of generation *n* take place on 31 December (a_2):

$$a_1 = \frac{D_1^0}{P_2^1 + D_1^0 + D_2^1} \quad \text{et} \quad a_2 = \frac{D_1^0}{P_1^0}$$

- For the lower triangle, the first hypothesis assumes that the migrations of year *n* of the generation born in *n-1* take place immediately after birth (b_1); the second, that the migrations of generation *n* take place exclusively on 31 December (b_2):

$$b_1 = \frac{D_2^0}{P_2^0 + D_2^0} \quad \text{et} \quad b_2 = \frac{D_2^0}{N_2}$$

We can thus calculate the upper and lower limits of $\hat{Q}$ by:

$$1 - \left[1 - \min(a_1, a_2)\right]\left[1 - \min(b_1, b_2)\right] \quad \text{et} \quad 1 - \left[1 - \max(a_1, a_2)\right]\left[1 - \max(b_1, b_2)\right]$$

These limits are of course close when the net migrations are low, both in the parallelogram with vertical base corresponding to generation *n-1* and in the lower triangle. We will use the mean of these limits.

The same procedure can be used to estimate the probability of dying between the 1st and 2nd birthday. The lower and upper limits are therefore:

$$1 - \left[1 - \min(a_1, a_2)\right]\left[1 - \min(b_1, b_2)\right] \quad \text{et} \quad 1 - \left[1 - \max(a_1, a_2)\right]\left[1 - \max(b_1, b_2)\right]$$

where

$$a_1 = \frac{D_1^1}{P_2^2 + D_1^1 + D_2^2}, \quad a_2 = \frac{D_1^1}{P_1^1},$$

$$b_1 = \frac{D_2^1}{P_2^1 + D_2^1} \quad \text{et} \quad b_2 = \frac{D_2^1}{P_1^0 - D_1^0}$$

***

It should be remembered that first marriage can be considered as producing repeatable or non-repeatable events. With regard to *longitudinal* indicators, the two processing methods yield the *same* results [28], but this is not the case for *transversal* indicators. In particular, the total rate based on the *rates* may well exceed the universe, which is never the case with a total rate based on *probabilities* or for the *longitudinal* indicator (proportion of ever-married persons at 50 years of age). Moreover, taking due account of the shift equal to the mean age at first marriage, the longitudinal and transversal indicators are generally much closer together when they are based on the probabilities than when they are based on the rates. Despite everything, the tradition observed in the publications of European statistics offices consists in treating first marriage in the same way as fertility, i.e. as producing repeatable events. However, the tools available to Eurostat will be able to treat first marriage *successively* as producing repeatable events, then non-repeatable events.

Sequence of operations

1 – Raw method (data in triangles or in squares)

- estimation of the rate: $t_i^n = \frac{C_i^n}{\left(P_i^n + P_i^{n+1}\right)/2}$

[28] This is why we often treat phenomena at non-repeatable events (first marriage or divorce) as if they produced repeatable events.

- conversion of the rate into a probability: ${}_1Q_i^n = 1 - e^{-t_i^n} \approx \frac{t_i^n}{1 + t_i^n/2}$

2- Advanced method (age completed)

a) constancy of the risk

- calculation of characteristics at birth: m_1 *and* m_2,
- calculation of the rate: $\hat{q}_i^n = \frac{E_i^n}{P_i^n m_1 + P_i^{n+1}(1 - m_2) + \frac{s_2 - s_1}{6}}$
- conversion of the risk into a probability: ${}_1Q_i^n = 1 - e^{-\hat{q}_i^n} \approx \frac{\hat{q}_i^n}{1 + \hat{q}_i^n/2}$

b) variable risk

- calculation of characteristics at birth: m and V, in their absence use values equal to $\frac{1}{2}$ and $\frac{1}{12}$ respectively.
- estimate of the instantaneous probability $\hat{q}_i^n$, calculated by successive *iterations* using the equation (3), assuming that:

  - either q is constant:

    at the 1st iteration: - we assume that $q_i' = 0$

    at each subsequent iteration: $q_i' = \frac{q_{i+1} - q_{i-1}}{2}$

  - or $\frac{q_i'}{q_i}$ is constant

    at the 1st iteration: - we assume that $\frac{q_i'}{q_i} = 0$

    at each subsequent iteration: $\frac{q_i'}{q_i} = 0{,}5 * \left(\frac{Log(q_{i+1})}{Log(q_{i-1})} \right)$

- estimation of the probability by:

$$ {}_1Q_i^n = 1 - e^{-q} \approx \frac{q}{1 + q/2} $$

3- Raw method (age reached)

- estimation of the rate: $t_i^n = \frac{C_i^n}{\left(P_i^n + P_{i+1}^{n+1}\right)/2}$
- conversion of the risk into a probability: ${}_1Q_{i-0,5}^n = 1 - e^{-t_i^n} \approx \frac{t_i^n}{1 + t_i^n/2}$

3- Advanced method (age reached)

a) variable risk

- calculation of characteristics at birth: m and V, in their absence use values equal to $\frac{1}{2}$ and $\frac{1}{12}$ respectively.
- estimation of the instantaneous probability $\hat{q}_i^n$, calculated by successive *iterations* using the equation (4), assuming that:

  - either q is constant:

    at the 1st iteration: - we assume that $q_i' = 0$

    at each subsequent iteration: $q_i' = \frac{q_{i+1} - q_{i-1}}{2}$

  - or $\frac{q_i'}{q_i}$ is constant

    at the 1st iteration: - we assume that $\frac{q_i'}{q_i} = 0$

    at each subsequent iteration: $\frac{q_i'}{q_i} = 0{,}5 * \left(\frac{Log(q_{i+1})}{Log(q_{i-1})} \right)$

- estimation of the probability by:

$$ {}_1Q_{i-0,5}^n = 1 - e^{-q} \approx \frac{q}{1 + q/2} $$

D) Fertility tables by order

Until relatively recently, living together without being married, the breakup of families due to divorce and remarriage were relatively uncommon. Thus, births occurred in almost all married couples, most often for the first time. Under these conditions, the definition used for the birth order had relatively little importance in terms of fertility indicators by order. The order defined in the current marriage differed little from that defined among all the births of the mother, the biological order[29]. Nowadays, with the development of new forms of union, resulting in a considerable increase in births outside wedlock, we can no longer analyze fertility according to the birth order with as much relevance if we have only the classification of births according to the order in the current marriage. Therefore, certain countries, such as France, have changed the definition and use instead the classification according to biological order, although this change was not easy: either survey data are used to produce indices that take account of all previous births (as is the case in France with the family survey) or an estimate is made more or less periodically of the biological order based on the order in the current marriage (such as in Germany) according to conversion keys based on a comparison of births, which are recorded during a survey and classified according to the two definitions.

A distinction should therefore be made between countries according to the definition of order used, and the indicators should be published in separate tables relating to one or other of the two definitions. Similarly, for countries that have recently changed the definition, we should ensure that the information based on different definitions is not mixed up, and the indicators should be presented in separate tables based on each of these definitions. If there is a choice, biological order should always take priority.

The data will be processed in the same way as the order is defined in the current marriage or among all the births of the mother. The only distinction that can be made, when the order is that defined in the marriage and therefore concerns only legitimate births, is that a column should be added for order *0* corresponding to births outside wedlock. In this way, the total of the thus completed table will be equal to the total number of live births, and its margin will give the distribution of births by age without distinction of birth order.

Similarly, we avoid having to make a calculation of the transversal and, more particularly, the longitudinal summary indices when the row is defined in the current marriage, as their interpretation is extremely complex.

Fertility by order, whether it is defined *within the current marriage* (distinguishing, among the live births based on the same union celebrated by marriage, the firstborn, second child, etc.) or *among all the births of the mother* (distinguishing, among the live births of the same woman, the firstborn, second child, etc.), by nature, produces *non-repeatable* events, but may be treated as producing *repeatable* events. This is particularly the case within the framework of the indices calculated by Eurostat.

The basic data available in the Eurostat database do not in fact provide a means of calculating the ratio of the births of each of the orders to women only who are subject to the risk of this event (the birth of a child of a specific order).

[29] The biological order of a live birth is the order of the child among the liveborn children of the *same mother*. Certain European countries do not observe the biological order in their routine civil status statistics, but only for births *in the marriage* (traditionally referred to as *legitimate* births) the order in the *current marriage*: order of the child among the liveborn children of the *same couple*. Fertility by order, which is defined by the order in the current marriage, is treated as producing repeatable events. Moreover, it cannot lead to a notion analogous to the parity progression ratio, at least if the frequency of births *outside* the *first marriage* (unmarried or remarried couples) is not zero.

Prior to the calculation of rates, quite often it is advisable to distribute the births by age and/or non-declared order using the *pro rata* procedure described in chapter IB1, page 23. This involves rectifying, when the two tables (by age and by age and order) are available, the non-declared elements based on the more detailed information given in the table by age and order and taking, for distribution by the rectified age of non-declared elements, the margin of the table by rectified order and age.

The fertility rates for each of the orders (biological or in the current marriage) will be calculated according to a procedure that is strictly identical to that used to calculate the general fertility rates by age. Only the events taken into consideration change: all the births of a given age, which appear in the numerator, are replaced by the only births of this given age of a given order.

The different possible calculation procedures remain the same, but we should of course use the same procedure in the two cases: general fertility by age and fertility by age and birth order.

In all rigour, when we carry out the same calculations as for general fertility we do not arrive at the same indices, even though a distinction is rarely made by the experts. In fact, we do not obtain rates for incidence or fertility by order as is often said, but rather *indices-components*. These different "rates" by order, for a given age, are in fact the result of the breakdown of the fertility rate by age into its different components by order, whereby the sum of the indices-components by order, for a given age, give the general fertility rate of this age. However, in the rest of this document the terms *component by age and order* and *rate by age and order* will be used interchangeably.

The sum of the components by age, of biological order *n*, for a given year, will give the component of order *n* of the total rate, sometimes referred to as the total rate of order *n*. We will discuss the component of order *n* of the completed fertility (or of the completed fertility of order *n* when the sum concerns the same generation). We will obtain the mean age - transversal or longitudinal - at childbearing of order *n* by applying the same calculation procedure as for the mean age at childbearing, all orders included.

The fertility by *biological order* of birth can also be treated as producing non-repeatable events. Thus, by considering the generations that have reached the *end* of their fertile life, we define the concept of the *parity progression ratios*[30]: the proportion, among women who have reached at a certain time in their life a given parity[31], of those who have at a later stage reached the directly superior parity. This indicator will be defined in chapter 3.

[30] This concept is useful for analysing the trend of fertility by order.
[31] The term *parity* refers in demography to the number of liveborn children.

E) Divorce rate tables

Divorce is generally treated as a repeatable event (which it is not really) occurring among *married couples*, and is studied differentially according to the *period elapsed since the celebration of the marriage* (also referred to as the *duration of marriage*). If there is no calculation system available, by marriage cohort[32], of the number of resident couples, which changes over the years due to mortality, divorce and international migrations, the divorce rates are determined by calculating the ratio of the numbers of events observed during the different periods to the *initial* total number of each marriage cohort. The data that are generally available correspond to the numbers of divorces according to the year of divorce and the duration elapsed since the marriage, *expressed in difference of year (millésime)*. The numbers of divorces observed each year are therefore classified by marriage *cohort*, i.e. by duration of marriage *reached* during the calendar year of the divorce (*parallelograms with vertical sides* of the Lexis diagram, with the exception of the divorces pronounced in the same year of the marriage, which correspond to a *triangle*).

The divorce rates, in contrast to the other rates mentioned above (fertility rate and first marriage rate), are treated as raw rates which do not involve the more or less irregular origin of marriage cohorts and do not therefore take into account the monthly trend of the number of marriages. This choice was made for two reasons:

- firstly, the number of surviving marriages is usually unknown except, perhaps, in certain countries with population registers or where the number has already been estimated from the mortaility of each sex and from migrations[33];
- secondly, all countries supply, as part of the joint collection process, divorces classified by the reached duration of the marriage in the year of divorce - a configuration in which measurement bias is always limited.

Yet this is no doubt fictitious. It is hard to believe that al countries record the year of marriage even though they do not record the year of birth of the mother, and that the table to be completed, supplied by the joint collection, does not give the Statistical Office a choice between the two possible definitions of duration and requires divorces to be classed by the duration reached. This is why, as a precaution and to guard against errors of interpretation of the definition of duration used, we do the following:

- ask each country for a precise definition of the duration it uses to classify divorces;
- consider the possibility of treating divorces, in the most appropriate way, according to every possible definition, i.e. taking into consideration variations in the input of marriage cohorts.

In this case we treat divorce like other demographic phenomena and calculate the *advanced* rates according to a transposed methodology of the calculation of the fertility rates. To this end, we must first calculate the characteristics, the year of marriage and the annual marriage cohorts, based on the monthly distribution of the total number of marriages, as is the case when we determine the characteristics, at birth, of the annual generations based on the monthly distribution of births. We then calculate the advanced divorce rates, as is the case for

[32] Only the distribution of resident persons by sex and year of age is the subject of regular calculation by the European national statistics offices.

[33] If we use marriage breakdowns by divorce when calculating surviving marriages, we find ourselves calculating occurrence-exposure rates (or probabilities) and no longer calculating additive incidence rates (reduced events).

fertility or first marriage, but replacing the denominator by the initial number of marriages in the mean population.

Raw rates may refer to the duration of marriage reached or the duration completed, depending on the case, but both use the *initial* number of marriages as the denominator. While the rate for duration reached *i* for the year of divorce *nn* is the ratio of the number of marriages celebrated in year *nn-i* and broken off in year *nn* to the total number of marriages celebrated in year *nn-i*, the rate for duration completed is the ratio of the number of marriages broken off in year nn to the duration *i* and the average number of marriages celebrated in years *nn-i* and *nn-i*-1.

As for any demographic phenomenon, we must set an interval of fixed ages at the occurrence of the events and therefore define a minimum age (or duration) and a maximum age, failing which the subsequent longitudinal recombination of the risk rates may be incorrect. Indeed, if the upper limit[34], for example, varies over time, we run the risk, insofar as in general this is an open interval, of overestimating the intensity of the phenomenon of cohorts for which we would have had to add rates corresponding to several consecutive open intervals.

The intensity of the divorce rate and the mean age of the marriage at the time of the divorce, whether they be transversal or in marriage cohorts, are calculated using the same procedures as in the case of other repeatable events.

When all the information required to calculate the divorce rates by annual duration of marriage is unavailable, it is, however, often possible to estimate the total divorce rate, as we shall see below. The method of estimation differs according to the type of information that is missing.

a) The initial total number of marriage cohorts corresponding to the highest durations of marriage is not available

In this case, we must admit, for want of anything better, the hypothesis according to which the annual number of marriages is, for these particular years, identical to that of the oldest year of which the initial cohort of marriages is known.

b) The classification of divorces by duration of marriage relates to multiannual groups of duration

The simplest procedure consists in calculating, for each multiannual group, the annual mean number of divorces by dividing the number of divorces of the group by its range (expressed in years) and then processing the annual series obtained by replacing the data for each duration by the appropriate annual mean number.

This situation is encountered quite frequently: the *terminal* duration thus corresponds to an *open* group of durations (for example, 30 years or more). In this case, the procedure to follow depends on the lower limit of duration to which this terminal group relates:

- If this lower limit is sufficiently high (for example, 35 years) to accept that the number of divorces is very low beyond this duration, we will proceed as if all the divorces of the open group had for common duration 35 years, and we will calculate the ratio of divorces at duration 35 years *or more* to the initial cohort of marriages of year n-35 only.
- If this lower limit is relatively high but not sufficiently high to accept the above approximation (terminal group: 25 years or more, for example), we can calculate the ratio of divorces at duration 25 years or more to the initial number of marriage cohorts that we consider to be the most affected by divorce at these durations.

[34] Although in the precise case of divorce it is the upper limit that should be limited, since the first duration is always duration 0, in respect of fertility and first marriage a minimum age must also be set.

We can also calculate the ratio of divorces at 25 years or more to the sum of the numbers of marriages celebrated from *n*-25 to *n*-30, if we consider that beyond 30 years of marriage there are no more divorces. As the rates thus calculated have an annual dimension, we should not forget to multiply them by the number of cohorts taken into consideration when we calculate the total divorce rate.

We can further divide the number of divorces of the open group by its assumed *effective* range, then calculate the ratio of each of the numbers obtained to the initial total number of each of the cohorts participating in this efficiency.

It can be intuited from this that the lower the lower limit of the terminal open group is[35], the more desirable it will be to weight each of the cohorts by its share of the risk of occurrence of divorce rather than to consider, as is the case in the solutions proposed above, that the risks are more or less equally distributed.

- If the duration is relatively low, the following three steps are taken:
    1. we calculate in the usual manner the divorce rates by duration until the last duration of annual amplitude;
    2. we calculate the ratio of the divorces of the open group to a weighted mean of the numbers of marriages leading to divorces at durations corresponding to the open group. The weighting coefficients are taken from a divorce calendar estimated or observed in another year or in another country.
    3. we obtain the total rate by summing the annual rates and adding the rate for the open group.

c) We have only the annual number of divorces and we do not have the distribution by duration of marriage.

It is possible to estimate the total divorce rate when the distribution of the number of divorces according to the duration of marriage is not available. The *method of the weighted mean* (also referred to as the *calendar type method*) yields this kind of estimate if we can refer to a plausible calendar of the occurrence of divorce over the durations of marriage.

This calendar, of which the sum of the elements equals 1, can correspond to an observation made, on another date in the country considered or in another country where the marriage and dissolution conditions are similar.

The effect of the choice of the calendar used on the level of the estimated indicator is slight if the number of marriages is not subject to sudden variations from one year to another.

Let us say that the ratio between the number of divorces D_x^n observed in year *n* at duration *x* and the initial number of marriages $M_g = M_{n-x}$ in the corresponding cohort[36] is equal to the product of the effective calendar element $a'_{g,x}$ of cohort *g* by the effective intensity p'_g of the divorce rate in this cohort, where:

$$a'_{g,x} = \frac{D_x^{g+x}}{\sum_{x=a}^{x=w} D_x^{g+x}} \text{ and } p'_g = \frac{\sum_{x=a}^{x=w} D_x^{g+x}}{M^g}$$

[35] Although it takes longer, a weighted mean is always preferable to a simple linear distribution. In the case of five-year groups of duration, it is always possible to distribute the number of events of the group among the cohorts concerned by referring to the share, observed in another year or in another country, of each of the rates by annual duration in the sum of rates by duration of this group.

[36] We will assume, to simplify matters, that the data are observed in parallelograms with vertical sides: therefore $g = n - x$.

The number of divorces D^n of a given year can therefore be written as the sum of the products of the *effective* calendar elements[37] $a'_{g,x}$ by the *effective* intensities p'_g of the divorce rate in the marriage cohort *g*, and by the total number of initial marriages M_g in these cohorts:

$$D^n = \sum_{x=a}^{x=w} a'_{g,x}\ p'_g\ M_g$$

The desired divorce rate indicator is:

$$ICD = \sum_{x=a}^{x=w} a_{g,x}\ p_g$$

which can be assimilated to:

$$\sum_{x=a}^{x=w} a'_{g,x}\ p'_g\ .$$

According to the hypothesis of stationarity of the divorce rate over time, the two above formulas can be written as follows:

$$D^n = p'_g \sum_{x=a}^{x=w} a'_{g,x}\ M_g$$

and:

$$S = p'$$

Therefore:

$$S = \frac{D^n}{\sum_{x=a}^{x=w} a'_{g,x}\ M_g}$$

This formula gives quite acceptable results when, on the one hand, the variations in the calendar and in the intensity are regular and moderate and, on the other hand, the disruptive events (migrations and widowhood) are not too marked. Indeed, this formula, in addition to the fact that it consists in substituting for the *effective* calendar and intensity of a *corrected* calendar and intensity the effects of the disruptive phenomena, presupposes that the divorce rate is stationary over time.

Moreover, in this method - as in the method used for the sum of the divorce rates by duration of marriage - the substitution of the *initial* total number of marriage cohorts with the cohort of subsisting marriages results in an underestimation of the total divorce rate if widowhood and emigration result in more marriages falling outside the scope of observation than immigration causes to fall within the scope of observation, and vice-versa.

Sequence of operations

1 – Raw rates (duration completed)

calculation of the rate in relation to the initial total number of marriages:

$$f_i^n = \frac{E_i^n}{(M^{n-i-1} + M^{n-i})/2}$$

2 – Raw rates (duration reached)

calculation of the rate in relation to the initial total number of marriages:

$$f_i^n = \frac{E_i^n}{M^{n-i}}$$

[37] i.e. taking account of the manifestation of disruptive events.

3- Advanced rates (duration completed)

a) constancy of the risk

- Calculation of the characteristics of marriage cohorts: *m_1 and m_2*
- Calculation of the rate:

$$\hat{f}_i^n = \frac{E_i^n}{M^{n-i-1}m_1 + M^{n-i}(1-m_2)}$$

b) variable risk

- calculation of characteristics of marriage cohorts: m, V and $\mathtt{m}$,; if these are missing, take values respectively equal to $\frac{1}{2}$, $\frac{1}{12}$ and 0.
- estimation of the rate, $f_i^n = f_i + \frac{f_i''}{24}$, by successive *iterations*

- at the first iteration: - assume that f_i' and f_i'' (first and second derivatives) are *zero* for all i

- using equation (1) we deduce f_i, then $f_i^n = f_i$, of E_i^n, M^{n-i-1} and M^{n-i}

- at each subsequent iteration : - we estimate f' and f'' by parabolic adjustment of the values for f obtained at the *previous* iteration

- again using equation (1), we deduce f_i, then $f_i^n = f_i + \frac{f_i''}{24}$, of E_i^n, M^{n-i-1} and M^{n-i}.

4- Advanced rates (duration reached)

- calculate the characteristics of marriage cohorts: m and V (if these are missing, take $\frac{1}{2}$ and $\frac{1}{12}$.
- Estimate the rate, $f_i^n = f_i + \frac{f_i''}{24}$, by successive *iterations*

- 1st iteration: - assume f_i' and f_i'' (first and second derivatives) to be *zero* for all i

- deduce, using equation (2), f_i, then $f_i^n = f_i$, of E_i^n, M^{n-i-1} and M^{n-i}.

- at each subsequent iteration: - estimate f' and f'' by parabolic adjustment of the values of f obtained at the *previous* iteration

- again using equation (2), we deduce, f_i, then $f_i^n = f_i + \frac{f_i''}{24}$, of E_i^n, M^{n-i-1} and M^{n-i}.

Chapter 3

Indicators deduced from the occurrence tables

A demographic table is a collection of *rates* or *probabilities*, according to age (or the duration elapsed since a date taken as the origin), relating either to the same *year of observation* (*transversal* table) or to the same *year of birth* (*longitudinal* table). From this table, we can deduce different summary indicators designed to *summarize* the information.

1- Derived indicators deduced from the tables

There are two types of indicators designed to summarize the information contained in a table: intensity indicator or central tendency indicator which summarizes the calendar of the phenomenon. We can therefore distinguish between the following:

- *Intensity* indicator in a table of *rates*: this is the *sum* of all the rates of the table. The indicator thus obtained bears the general name *total rate* of the phenomenon considered if the table is *transversal*. If the table is *longitudinal*, it bears a different name according to the phenomenon: *completed fertility* in the case of *fertility*, *proportion of ever-married persons at 50 years* in the case of *first marriage*, *proportion of marriages dissolved by divorce* in the case of the *divorce rate*.

  As we have seen in the previous chapter on the divorce rate table, in the absence of elements that provide a means of calculating a transversal table and thus of deducing the summary intensity indicator, it is very often possible to produce an estimation of this. These estimation procedures are all based on the method of the weighted mean used by Corrado Gini. It is this method that we have used to propose an estimation of the total divorce rate in the absence of the distribution of divorces according to age of marriage. It is on this same principle that Gérard Calot based his concept of *mean generation* (to which we will return) which provides a means of calculating several months before the elements required for this calculation are available an estimation of the total rate based only on the availability of the absolute number of events[38].

- *Intensity* indicator of a table of *probabilities*: this is the proportion of statistical units that experience the event considered between the extreme ages of a and w. The indicator thus obtained bears the *same* name whether the table is transversal or longitudinal, but we then specify which, for example, the *proportion of ever-married persons at 50 years* based on the transversal or longitudinal *probabilities* in the case of first marriage. With regard to mortality (fatal event), the intensity indicator (transversal or longitudinal) is equal to 1 and is not applicable[39].

- *Central tendency* indicator of a table of *rates*: this is, when the event studied occurs, the mean age of persons concerned at the time of its occurrence. The

[38] cf. G. Calot, "Une notion intéressante: l'effectif moyen des generations soumises au risque". I. Présentation méthodologique", *Population*, 6, 1984, 947-976.

[39] It should, however, be mentioned that, as is commonly the case for the analysis of first marriage, mortality (*transversal*) could also be treated as a repeatable event. In this case, we calculate the incidence rates, which can be assimilated to the events of a table, which entail the calculation, by summation of these rates, of a total mortality rate (of which the value is not 1) and of a mean age at death. For further details see "Un indicateur conjoncturel de mortalité : l'exemple de la France", *Population*, n° 2, 1993, pp. 347-368.

indicator thus obtained bears the name *mean age*, whether the table is *transversal* or *longitudinal*: *mean age at childbearing* in the case of *fertility*, *mean age at first marriage* based on the rates (single persons marrying before the age of 50) in the case of *first marriage*, *mean age of marriages dissolved by divorce* in the case of the *divorce rate*. The mean ages which constitute the central tendency indicators of repeatable events (fertility) or are treated as such (first marriage) must therefore *be based on the rates* and not on the events.

The formula for the calculation of this mean age varies according to the figure of the Lexis diagram in which the rates used are calculated. Thus, when the rates are by age completed or between birthdays (Lexis square), the mean age is calculated as follows:

$$\overline{x} = 0{,}5 + \frac{\sum_{x=a}^{x=w} x f_x^n}{\sum_{x=a}^{x=w} f_x^n}$$

If the rates are defined according to age reached during the year or in difference of year (*millésime*) (parallelogram with vertical base of the Lexis diagram), the formula becomes:

$$\overline{x} = \frac{\sum_{x=a}^{x=w} x f_x^n}{\sum_{x=a}^{x=w} f_x^n}$$

where *n* is the calendar year when we are concerned with the mean transversal age, and the cohort (year of birth or of marriage) within the framework of the longitudinal analysis.

- *Central tendency* indicator of a table of *probabilities*: this is again, when the event studied occurs, the mean age of the persons concerned at the time of its occurrence. The indicator thus obtained bears the name *mean age at death* or *life expectancy at birth* in the case of *mortality* and *mean age at first marriage* based on the *probabilities* (single persons marrying before the age of 50) in the case of *first marriage*.

  The formula for the calculation of the mean age at death or life expectancy at birth is very close to that of the mean age in the case of a repeatable event identified according to age in completed years, the only difference being that we use the events of the table $d_{x,x+1}$[40]:

$$e_0 = 0{,}5 + \frac{\sum_{x=a}^{x=w-1} x * d_{x,x+1}}{\sum_{x=a}^{x=w-1} d_{x,x+1}}$$

  which is simplified as:

[40] To which the incidence rates can be assimilated.

$$e_0 = 0{,}5 + \frac{\sum_{x=1}^{x=W} S_x}{S_0}$$

where S_x is the number of survivors, at age x, of the life table.

The following can be added to these conventional indicators:

- *Definitive infertility*[41]: this is the one's complement of the intensity indicator of the table of fertility rates of biological order 1, in which the first birth order - non-repeatable event - is treated as a repeatable event by the calculation of incidence rates which, in all rigour, are indices-components and yield the component of order 1 of the different rates of general fertility by age. It is calculated transversally based on the total first birth order and longitudinally based on the completed fertility of order. This complement of the intensity indicator is also frequently used in the analysis of the first marriage, which is known by the name *frequency of definitive celibacy*.

- The *proportion of women with x children*: this is the proportion of women who have given birth to *x* children among all women. It can be calculated transversally as in cohorts. The proportion of women *with no* children - measurement of infertility - is, transversally, the one's complement of the total first order rate; in the generations this is the one's complement of the completed fertility. For the other dimensions of the family, it is the difference between the intensity indicator of fertility of biological order *n* and that of order *n+1*. For the last dimension used, the proportion of women who have had *4 children or more*, for example, is equal to the intensity indicator of the fertility order used, the total rate or the completed fertility of order *4 or more*, depending on the case.

- The parity progression ratios a_n are special derived indicators of a collection of fertility tables by biological order. Thus, the parity progression ratio of order 1 to order 2 (expressed also as a_1) is the proportion among women who have had *at least one child* of those who have had *at least two*. Similarly, the parity progression ratio of order 0 to order 1 (expressed also as a_0) is the proportion among *all* women of those who have had *at least one* child (its unit's complement is the *proportion of women with no children*, also referred to as the *frequency of definitive infertility*). The parity progression ratios are defined in *transversal* terms: the proportions of women who have reached the different parities are obtained on the basis of the cumulations of all ages of the transversal fertility rate by age, as are the analogous longitudinal proportions. Eurostat intends to calculate the longitudinal and transversal parity progression ratios *successively* in its system.
  The longitudinal parity progression ratios can be calculated on as yet *incomplete* paths of fertility. Thus, at age *i*, we will calculate the proportion[42] among women of age *i* who have already had *at least r* children of those who have already had *at least r*+1. Each proportion is obtained by calculating the ratio of the intensity indicator of order *n+1* to the intensity indicator of order *n.*
  The combination of these ratios, according to the formula:

$$a_0 + a_0\, a_1 + a_0\, a_1\, a_2 + \ldots$$

[41] To be relevant, this indicator must be calculated using one single classification of live births according to the biological order of the birth, i.e. among all the live births of the mother.

[42] The proportion among women of 50 years of age who have had *at least r* children of those who have had *at least r*+1 is precisely the longitudinal parity progression ratio $a_{r \dashrightarrow r+1}$.

once again gives the intensity indicator of general fertility, the total rate or the completed fertility, according to the case.

2- Precautions to be taken when calculating these indicators

In the previous paragraph, it was mentioned that the intensity indicator of a table of rates is the sum of all the rates of the table. Thus, the total fertility rate, for example, is the sum of all the fertility rates by age available for a given year.

However, several difficulties may arise when the interval of available ages varies over time:

- Firstly, the *transversal* indicator of central value, the mean age at childbearing, for example, which is, by construction, rather sensitive to extreme values, may reflect these variations in limit ages of the tables;
- Secondly, and this is undoubtedly the most important element, the longitudinal transposition of the rates may lead to measurements of the longitudinal intensity of which the calculation procedure would vary from one cohort to another.

This variability in limits would have no incidence on the quality of the measurements if it involved only the random variations in the expression of the studied phenomenon, but in most cases the statistics offices make groupings at both extremities of the ages. If these open intervals are not always identical, in the longitudinal recombination we run the risk of assigning to the same cohort several observations referring to the same age and thus of overestimating its intensity.

Therefore, if in a given year *n* the last open group is 49 years or over and in the next year *n+1* it increases to 50 years or over, we will assign to the generation born in *n-49* the two open groups (the 49+ group of year n and the 50+ group of year n+*1*). Therefore, ages 50 et seq. will be counted twice.

It is for this reason that an unvarying age interval that is fixed for each of the phenomena must be in the table calculation program. This could be 15-49 years for all the manifestations of fertility and nuptiality, 0-30 years for the divorce rate and 0-99 years for mortality.

3- Other derived indicators

The statistical data observed in the same year or in the same generation can be combined in many ways to arrive at derived indicators.

a) Crude rates

A *crude rate* is a ratio between an *absolute annual number of events* and the mean cohort of the resident population[43]: *crude birth rate* (absolute annual number of *live births*), *crude death rate* (absolute annual number of *deaths*), *crude marriage rate* (absolute annual number of *marriages*), crude divorce rate (absolute annual number of *divorces*); the ***relative net migration*** is also a crude rate.

The mean cohort of the resident population that should be used is the *half-sum* of cohorts of the population at the start and end of the year. However, certain European countries which have a population register (from which they extract *quarterly* and even *monthly* data) take as the mean population of the year the arithmetic mean of quarterly or monthly cohorts. We feel that it would be eminently desirable to convince them to do away

[43] More precisely, but equivalent if the natural change and net migrations do not have a seasonal movement, the denominator is the number of person-years of exposure to the risk.

with this refinement, given the very rough nature of a crude rate as a demographic indicator and the very slight differences that generally separate the crude rates obtained.

b) Mean cohort of generations subject to the risk

In the case of a *repeatable* phenomenon, the absolute number of events observed in a given year is the product of the *total rate* by the *mean cohort of generations subject to the risk*. Thus, with regard to fertility, this mean cohort is the weighted mean of the cohorts of the different female generations which, in that year, are of fertile age (15-49 years, to be precise), whereby the weighting coefficient of the cohort of a given age is the fertility rate at that age[44].

This indicator is a measurement of the *size* of the country with regard to the phenomenon under consideration. It is characterized by *slow* variations, as a moving mean, which facilitates its interpolation and especially its extrapolation. Thanks to its extrapolation, we can estimate the total rate once we have an estimate of the absolute number of events, without waiting to have cohorts by age and thus the corresponding table.

It is desirable to have, for each country and each repeatable event (fertility) or event treated as such (first marriage, divorce), this derived indicator, extrapolated over a few years. It would then be up to Eurostat to determine whether this indicator should be published.

c) Longitudinal recombination of rates and probabilities: longitudinal indicators

The annual collections of rates and probabilities lead to the corresponding longitudinal collections after reclassification according to year of birth (or couple of years of birth in the case of rates and probabilities calculated using data in squares in the Lexis diagram). These longitudinal collections are summarized using derived indicators analogous to the derived indicators of transversal collections.

One question which arises in the longitudinal case and which does not have a transversal equivalent is the following: if a generation that has almost reached the *end* of its period of exposure to the risk still does not have the *last* rates or probabilities, we cannot summarize the behaviour of this generation by means of the usual derived indicators. However, the last rates or probabilities are often *small*, and their estimation would provide a means of supplementing the collection and estimating without any major risk of error the corresponding derived indicator.

Several methods can be used to make this estimate.

The first is the *freeze* method: the missing rate or probability is estimated, whatever the generation considered, on the basis of the most *recently* observed rate or probability at the *same age* (or at the same duration). Of course, the use of this method should be *limited*: for example, if we estimate the completed fertility, we will confine ourselves to the generations of which the sum of the estimated missing rates does not exceed 15% of the estimated completed fertility. In addition, it would be desirable to give the estimation of the completed fertility an *indicative letter* (for example, * if less than 1% has been estimated, # if between 1% and 5% has been estimated, etc.). The use of an indicative letter is also recommended when we use one or other of the methods below.

The second method involves the *extrapolation* of the missing rates and probabilities. This extrapolation can be carried out on a longitudinal basis or on a transversal basis.

- For example, on a longitudinal basis the fertility rates of a given generation are known up to age *i*-1: the missing part of the completed fertility is the sum of the

[44] For further details, see appendix, *L'analyse démographique conjoncturelle*.

rates of age i at age w. We consider the generations *prior to* the generation considered of which we know the *entire* history, and we extrapolate the series of partial fertilities at i years or more by a given process (for example, *linear* adjustment over the *ten* most recent generations of which we know the entire history).

- For example, on a transversal basis the fertility rate not yet observed of a given generation is estimated at a given age by extrapolation of the rates observed at that same age over the years. The estimates thus obtained are added together to evaluate the sum of the rates until age w of the generations that have not yet reached the end of their fertility.

A third method, which is applicable only to longitudinal indicators of *intensity* based on *rates* (repeatable events or events treated as such), is as follows: the longitudinal sum $\Sigma(n,i)$ of the missing rates of generation *g* from age *i* until age w is approximately equal to the *transversal* sum $S(n,i)$ of the rates between these same ages for the year of observation *n* when $n = g + AMT(n, i)$, where $AMT(n, i)$ is the mean transversal age at the occurrence for the events that occur in year *n* at *i* years or over[45]. It can be demonstrated that, as a whole, this third method is markedly more accurate than the first two methods.

d) The construction of monthly derived indicators

Knowledge of the absolute numbers of events on a *monthly* scale, all ages included, and not only on an *annual* scale, provides a means of calculating total monthly rates for fertility, first marriage by sex and mortality by sex. These monthly indicators describe more *precisely* the chronological trend of the phenomena studied than the annual indicators, and prove, where appropriate, the simultaneity (and thus the probable causal link) between an anomaly in the trend of the indicator and the occurrence of such or such an event (political, social or other). The methodology of the construction of monthly indicators by the EDO was presented in the article (which is reproduced in the annex) by Gérard Calot entitled "*L'analyse démographique conjoncturelle*", which was a contribution to the volume *The joy of demography* published in honour of Dirk J. van de Kaa, by Anton Kuijsten, Henk de Gans and Henk de Feijter, NethurD Publications, The Hague, 1999.

e) Short-term extrapolation of annual transversal indicators

When, for a given country, the data relating to a given (recent) year are not yet available, it is not possible to produce an indicator for that country in that particular year or to estimate anything at the level of any geographic aggregation containing said country. For this reason, especially if the population of the country in question has a relatively low cohort in relation to the cohort of the aggregated zone, it is desirable to estimate, even roughly, the missing data.

The population cohorts will be estimated by assuming the invariance, given equal ages, of the apparent survival ratios; the absolute numbers of events by age will be estimated by assuming the invariance, given equal ages, of the rates, whereby these two invariances are assessed in relation to the most recent year for which the information is available.

If, in addition, we have an estimate of the annual number of events, we will correct proportionally the annual numbers of events by age to make their total coincide with the estimated annual number.

With these roughly estimated data, we proceed as if they were observed data.

[45] Applied to $i = \bullet$, this equation gives only very approximate results: it consists, for example with regard to fertility, in assimilating the total rate to the completed fertility, using the transversal shift of the mean age at childbirth. This equation becomes steadily more precise as *i* approaches $\bullet$.

Formulation of derived indicators			
	Type of age	Repeatable event	Non-repeatable event
$\hat{L}\left(a+\frac{1}{2}\right)$	Reached	///	$\hat{L}\left(a+\frac{1}{2}\right)=1-\prod_{i=\mathrm{a}}^{i=a}\left(1-p_i^{G+i}\right)$
$\hat{L}(a)$	Completed	///	$\hat{L}(a)=1-\prod_{i=\mathrm{a}}^{i=a-1}\left(1-p_i^{G+i}\right)$
$\hat{D}(G)$	Reached or completed	$\sum_{i=\mathrm{a}}^{i=\mathrm{w}} p_i^{G+i}$	$1-\prod_{i=\mathrm{a}}^{i=\mathrm{w}}\left(1-p_i^{G+i}\right)$
$\bar{x}(G)$	Reached	$\frac{\sum_{i=\mathrm{a}}^{i=\mathrm{w}} i\ p_i^{G+i}}{\sum_{i=\mathrm{a}}^{i=\mathrm{w}} p_i^{G+i}}$	$\frac{\sum_{a=\mathrm{a}}^{a=\mathrm{w}} a\left[\hat{L}\left(a-\frac{1}{2}\right)-\hat{L}\left(a+\frac{1}{2}\right)\right]}{1-\hat{L}(\mathrm{w})}$
	Completed	$\frac{\sum_{i=\mathrm{a}}^{i=\mathrm{w}}\left(i+\frac{1}{2}\right) p_i^{G+i}}{\sum_{i=\mathrm{a}}^{i=\mathrm{w}} p_i^{G+i}}=\frac{1}{2}+\frac{\sum_{i=\mathrm{a}}^{i=\mathrm{w}} i\ p_i^{G+i}}{\sum_{i=\mathrm{a}}^{i=\mathrm{w}} p_i^{G+i}}$	$\frac{1}{2}+\frac{\sum_{a=\mathrm{a}}^{a=\mathrm{w}-1} a\left[\hat{L}(a-1)-\hat{L}(a+1)\right]}{1-\hat{L}(\mathrm{w})}$
$\hat{D}(a, b)$	Reached or completed	$\sum_{i=a}^{i=b-1} p_i^{G+i}$	$1-\prod_{i=a}^{i=b-1}\left(1-p_i^{G+i}\right)$

Where $\hat{L}(a)$ = probability that the event will not occur before the exact age *a*,

$\hat{D}(G)$ = final intensity of the occurrence of the event,

$\bar{x}(G)$ = mean age at the occurrence of the event,

$\hat{D}(a, b)$ = intensity of the occurrence of the event between the exact ages *a* and *b*.

Conclusion

The definition of the methodology to be followed with regard to the calculation of demographic indicators is of course a crucial aspect of the organisation of an international system for calculating demographic indicators, as without an appropriate methodology it is not possible to obtain good comparability of data.

Certain proposals may appear superfluous or over-complex for a rarely occurring problem, but it should be borne in mind that a system of this kind is not upgraded regularly and that it must be able to deal with all situations. Of course, quality comes at a price, but if we are to ensure the perennity of such a system, there are no shortcuts to quality.

However, although a good methodology is one condition for an efficient system, it is not a sufficient condition. There are many other aspects that can be just as important to guarantee the success of such an undertaking.

Therefore, we must:

- use tools to control the quality and internal coherence of the data incorporated in the database, which will involve regular purges;
- create the conditions which provide a means of processing in a unique manner all the available information, irrespective of its structure (this point was discussed in chapter one);
- design the articulation among all the modules thus defined in order to develop a coherent and complete system which continuously goes through all the stages from incorporation of the databases to the production of derived indicators;
- define a sufficiently flexible structure to permit the easy incorporation of all the additional modules required to meet new needs that will certainly be perceived in future.

ANNEX 1

The construction of life tables

Reproduit de :Gérard Calot et Ana Franco

“The construction of life tables”

in Guillaume Wunsch, Michel Mouchart et Josianne Duchêne (éditeurs),

Life Tables : Data, Methods, Models, Kluwer, 2001,

pp. 31-75

THE CONSTRUCTION OF LIFE TABLES

by Gérard CALOT[46] and Ana FRANCO[47]

The paper first recalls the probabilistic background of the estimation of mortality intensity by age and sex during a year of observation. It leads to the fundamental pair of relationships – hereafter numbered (24) and (32) – yielding the number of deaths inside each triangle of the Lexis diagram.

Three methods (the last one with two additional sub-variants) of practical computation of the probability of dying between two consecutive ages are considered : the first one (*A*) is based on the number of deaths inside each *triangle*, taking into account the distribution of *birthdays* within each of the two birth-cohorts concerned, the second one (*B*) is based on the number of deaths inside each *square* and the distribution of birthdays, the third one (*C*) simply consists in the computation of *rates* : ratio of the number of deaths in the square to the *mid-year* population. Method *C* takes the *observed rate* as an estimate of the *risk of mortality q*, while method C_2 takes it as an estimate of the *probability of dying Q*. Method C_1 is intermediate between *C* and C_2[48].

It is shown that, except for ages 0 and 1 and for higher ages (above 80), and except for cohorts born in a period of *abrupt* changes in the birth-rate, method *C*, and even methods C_1 and C_2, yield satisfactory results. But for cohorts born during world wars, it is really worthwhile to use methods *A* or *B* and – at higher ages – to use method *A*. Moreover if methods *C* and C_1 yield quite comparable results, method C_2 becomes inaccurate above age 60.

I – The estimation of a constant risk

Let us consider a set of *N* individuals exposed to the occurrence of a *non repeatable* event, denoted *E*. Individual *j*, $j = 1, 2, ..., N$, is exposed to *E* between times b_j and e_j, beginning and end of the *exposure* period, respectively.

Let us assume that the *intensity* of occurrence of *E*, denoted *q*, also called the *risk*[49] or the *instantaneous quotient* of occurrence of *E*, is *constant* over time and *identical* for all individuals.

Under these assumptions, the probability that event *E* occurs to individual *j* between times *t* and $t + dt$, given that it did not occur to that individual before time *t*, is, for any *j* and *t* :

---

[46] Observatoire Démographique Européen, Saint-Germain-en-Laye, France

[47] Eurostat, Statistical Office of the European Communities, Luxembourg.

[48] Methods *C*, C_1 and C_2 consist in deriving the estimate $\hat{Q}$ of the probability of dying, *Q*, from the observed *rate r* :

$$r = \frac{2D}{P_1 + P_2}$$

by :

$$\text{Method } C : \hat{Q} = 1 - e^{-r}$$

$$\text{Method } C_1 : \hat{Q} = \frac{r}{1 + r/2}$$

$$\text{Method } C_2 : \hat{Q} = r$$

[49] If event *E* is *death*, *q* is called the *force of mortality*.

$$q\,\mathrm{d}t. \qquad (1)$$

Let $P(b, t)$ denote the probability that event E does not occur to a given individual exposed to E between times b and t. We may write :

$$P(b,\ t\ +\ \mathrm{d}t)\ =\ P(b,\ t)\ (1\ -\ q\,\mathrm{d}t), \qquad (2)$$

expressing that the probability $P(b,\ t\ +\ \mathrm{d}t)$ of no occurrence between times b and $t + \mathrm{d}t$ is :

- the probability $P(b,\ t)$ of no occurrence between times b and t multiplied by
- the probability $1\ -\ q\,\mathrm{d}t$ of no occurrence between times t and $t + \mathrm{d}t$, given its non-occurrence between times b and t.

From (2), it follows that P satisfies :

$$\frac{\partial\left\{\mathrm{Log}\left[P(b,\ t)\right]\right\}}{\partial t}\ =\ -q, \qquad (3)$$

which implies, taking into account that $P(b, b) = 1$, that :

$$P(b\ ,\ t)\ =\ \mathrm{e}^{-q\,(t\,-\,b)}. \qquad (4)$$

Therefore, the probability that E occurs to individual j between times t and $t + \mathrm{d}t$ is :

$$\mathrm{e}^{-q\left(t\,-\,b_j\right)}\ q\,\mathrm{d}t. \qquad (5)$$

The probability of occurrence of E during a *unitary* period – i.e. of length equal to 1 – is denoted Q. Using (4), the probability of no occurrence during a unitary period, $1\ -\ Q$, is :

$$1\ -\ Q\ =\ \mathrm{e}^{-q},$$

so that Q is related to q by :

$$Q\ =\ 1\ -\ \mathrm{e}^{-q}\ \approx\ \frac{q}{1+\frac{q}{2}}\quad \text{if } q \text{ is small.} \qquad (6)$$

Let us now consider a set of N individuals, *independently* exposed to the occurrence of event E. The probability that E occurs to individuals $j_1, j_2, \ldots, j_n$ at times $t_{j_1}, t_{j_2}, \ldots, t_{j_n}$ and *not* to the other $N\ -\ n$ individuals, is :

$$\mathrm{L}\ =\ \prod_{j=j_1}^{j_n}\left[\mathrm{e}^{-q\left(t_j\,-\,b_j\right)}\,q\right]\ \prod_{j\,\notin\,(j_1,\,\ldots,\,j_n)}\left[\mathrm{e}^{-q\left(e_j\,-\,b_j\right)}\right]. \qquad (7)$$

Let T_j be, for individual j, the time elapsed from the beginning b_j of the exposure period until :

- the time of occurrence of E, if E did occur to individual j ;
- the end e_j of the exposure period, if E did not occur to individual j.

T_j is called the *duration of the period at risk*[50] for individual j. Likelihood L thus satisfies :

$$\frac{\partial\left[\mathrm{Log}(\mathrm{L})\right]}{\partial q}\ =\ n\ \mathrm{Log}(q)\ -\ q\ \sum_{j\,=\,1}^{N} T_j. \qquad (8)$$

[50] The *exposure period* of individual j is (b_j, e_j), but his *period at risk* is (b_j, e_j) if event E does not occur to him, (b_j, t_j) if event E occurs to him at time t_j.

The maximum likelihood estimate of q, derived from the sample, is therefore :

$$\hat{q} = \frac{n}{\sum_{j=1}^{N} T_j} \qquad (9)$$

Thus, the risk, when *constant* over *time* and *identical* among *individuals*, is estimated, on the basis of a sample of *independent* exposures, by the ratio of the *number of events recorded* to the *sum of the durations of the periods at risk.*

If the available information results from *several* independent samples satisfying the same assumptions with identical risk, the overall estimate of this common risk is the ratio of the *total* number of events recorded to the *total* sum of the durations of the periods at risk, *total* meaning computed on the *sum* of the various samples.

It then follows that the overall estimate of q is the *weighted harmonic mean* of the estimates yielded by the different samples, the weights being the *numbers of events* recorded :

$$\hat{q} = \frac{\sum_{s=1}^{m} n_s}{\sum_{s=1}^{m} \left(\sum_{j=1}^{N_s} T_{j,s} \right)} = \frac{\sum_{s=1}^{m} n_s}{\sum_{s=1}^{m} \left(\frac{n_s}{\hat{q}_s} \right)} \qquad (10)$$

If the N individuals are exposed to E during periods of *unitary* duration i.e. $\forall j : e_j - b_j = 1$, then the maximum likelihood estimate $\hat{q}$ of q differs from the *proportion* f of individuals recording event E. In fact, we have :

$$f = \frac{n}{N} \quad \text{with} \quad E(f) = 1 - e^{-q} = Q ,$$

which shows that f is an unbiased estimate of Q, while :

$$\hat{q} = \frac{n}{\sum_{j=1}^{N} T_j} = \frac{n}{N \overline{T}} = \frac{f}{\overline{T}}$$

is a slightly biased[51] estimate of q :

$$E(\hat{q}) = E\left(\frac{f}{\overline{T}} \right) \neq \frac{E(f)}{E(\overline{T})} = q$$

since :

$$E(\overline{T}) = E(T) = \int_0^1 e^{-qx} \, qx \, dx + e^{-q} = \frac{1 - e^{-q}}{q} = \frac{E(f)}{q}$$

[51] It can be shown that, for large N, the mathematical expectation of $\hat{q}$ is :

$$E(\hat{q}) \approx q \left[1 + \frac{1}{N} \frac{1 - e^{-q}(1 + q)}{(1 - e^{-q})^2} \right] \approx q \left(1 + \frac{1}{2N} \right) \quad \text{if } q \text{ is small.}$$

Considering a sample of N individuals, if *only* the number n of occurrences of E is known, i.e. if n is known, but not $\overline{T}$, we can base an estimate of q on f. We have :

$$E(f) = 1 - e^{-q},$$

that is :

$$q = -\mathrm{Log}[1 - E(f)].$$

Since f converges in probability (towards $1 - e^{-q}$), we can base the following estimates of q and Q on f :

$$\begin{aligned} \hat{q} &= -\mathrm{Log}(1-f) \approx \frac{f}{1-\frac{f}{2}} \quad \text{if } f \text{ is small} \\ \hat{Q} &= 1 - e^{-\hat{q}} = f \end{aligned} \tag{11}$$

An equivalent means of computing $\hat{q} = -\mathrm{Log}(1-f)$ is derived as the limit of the convergent sequence defined by the iterative relationship :

$$q_{k+1} = f \frac{q_k}{1 - e^{-q_k}} \tag{12}$$

starting from the limit of the right hand side of (12) when $q \to 0$:

$$q_0 = f . \tag{13}$$

As an application of these results, let us consider the estimation of the probability of dying between exact ages i and $i + 1$ on a *closed* cohort – i.e. *without* migration – (Figure 1).

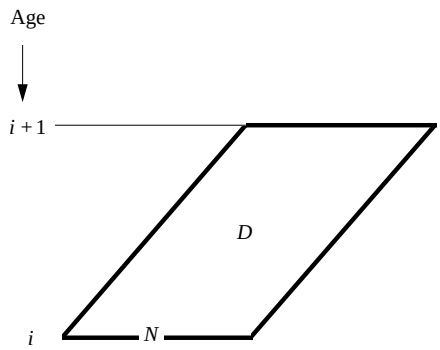


Figure 1. Estimation of the probability of dying in a parallelogram with horizontal sides.

The duration of the exposure period, between birthdays i and $i + 1$, is one year for *each* individual. The risk – also called the *force* – of mortality, q, is assumed to be *constant* with age – at least when age ranges between i and $i + 1$ – and to have no *seasonal* variation : q is the same for all individuals during the *whole* of their period at risk, whatever the location of their birthday within the year. If D is the number of deaths recorded and N the number of members of the cohort who were present at their i^{th} birthday, then q and Q are respectively estimated by :

$$\begin{aligned} \hat{q} &= -\mathrm{Log}\left(1 - \frac{D}{N}\right) \approx \frac{\frac{D}{N}}{1-\frac{D}{2N}} \quad \text{if } \frac{D}{N} \text{ is small} \\ \hat{Q} &= 1 - e^{-\hat{q}} = \frac{D}{N} \end{aligned}$$

II. – Variable risks

If we drop the assumption that q is *constant* over time, relation (4) yielding the probability that event E does not occur between times b and t, becomes :

$$P(b,\ t) \ = \ \mathrm{e}^{-\int_b^t q(x)\,dx} \qquad (14)$$

and quotient Q_i, between exact ages i and $i + 1$, i.e. in a unitary period, is given by :

$$Q \ = \ 1 \ - \ \mathrm{e}^{-\int_0^1 q(x)\,dx} \qquad (15)$$

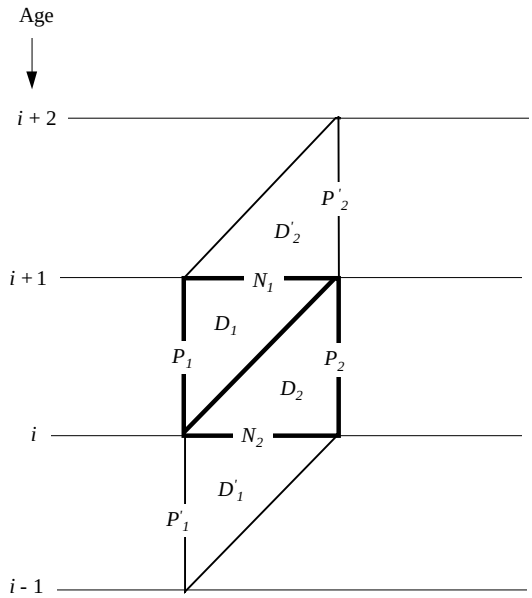
The estimation problem of q and Q, when exposure periods differ among individuals, remains parametrical only if we specify the form of risk function $q(t)$. In the following developments, we shall assume that risk q varies *linearly*[52] with time, with a derivative q' remaining *small* compared to q, and we shall consider the case of *mortality* within triangles of the Lexis diagram. Ages are denoted i and the calendar year under consideration is T.

In this specific case, it is assumed that, except for the beginning of life (i.e. for ages $i = 0$ or $i = 1$), the *force of mortality* $q(y)$ at *exact* age y, $i < y < i + 1$, for persons belonging to the infinitesimal cohort born between g and $g + \mathrm{d}g$, is a *linear* function of y but does not depend on g. Mortality is thus free from *seasonal* variations.

Furthermore, the following assumptions are made concerning migrations :

(i) migrations are *uniformly* distributed inside each triangle. More precisely, the *net migratory balance* between ages y and $y + \mathrm{d}y$, among the infinitesimal cohort born between g and $g + \mathrm{d}g$, is assumed to be $s\,\mathrm{d}g\,\mathrm{d}y$, where s is *constant* inside a parallelogram with vertical sides, made up of two adjacent triangles (Figure 2).

(ii) immigrants, from the time of their arrival, and emigrants, until the time of their departure, are exposed to the *same* risk of death as the non migrants of the same age.



[52] An alternative assumption, in line with Gompertz's law, which is generally close to reality at adult ages above 30 or 40 years of age, consists in assuming that it is not q' but q'/q, i.e. the *logarithmic* derivative of q, that is constant.

Figure 2. Notation of population numbers at year ends and on birthdays and notation of numbers of deaths

These two assumptions regarding migrations are obviously *simplistic*. They enable to take an extremely *rough* account of mobility. But if, at all ages, the magnitude of migrations is small compared to the size of the resident population, the impact on mortality is necessarily small and the correction for migrations derived from these assumptions is to be viewed as a *crude* correction for the *volume* of net migrations. In fact, alternative assumptions concerning the seasonal pattern of net migrations or differential mortality between migrants and non-migrants would be very difficult to formulate.

Inside the upper parallelogram of Figure 2, the value of s can be derived from the populations present at the end of each year and recorded deaths in the relevant triangles :

$$s_1 = P'_2 - P_1 + D_1 + D'_2 \qquad (16)$$

and the number of $(i + 1)^{th}$ birthdays celebrated during year T is :

$$\begin{aligned} N_1 &= P_1 - D_1 + \frac{s_1}{2} \\ &= \frac{P_1 + P'_2}{2} + \frac{D'_2 - D_1}{2} \end{aligned} \qquad (17)$$

Similarly, inside the lower parallelogram, the value of s is :

$$s_2 = P_2 - P'_1 + D'_1 + D_2 \qquad (18)$$

and the number of i^{th} birthdays celebrated during year T is :

$$\begin{aligned} N_2 &= P_2 + D_2 - \frac{s_2}{2} \\ &= \frac{P_2 + P'_1}{2} + \frac{D_2 - D'_1}{2} \end{aligned} \qquad (19)$$

In each square of the Lexis diagram, two sorts of triangles will be considered : the *upper* triangle and the *lower* triangle. Let us start with the lower one (Figure 3).

Relationship in the lower triangle

Let $\mathrm{d}N_2 = N_2\, g_2(u)\, \mathrm{d}u$ denote the number of persons celebrating their i i^{th} birthday between times u and $u + \mathrm{d}u$, time 0 being January 1st of year T. The *density* of the distribution of life lines at time u and birthday i is thus $g_2(u)$, among the N_2 members of the annual cohort born during year $T - i$, present at their i^{th} birthday.

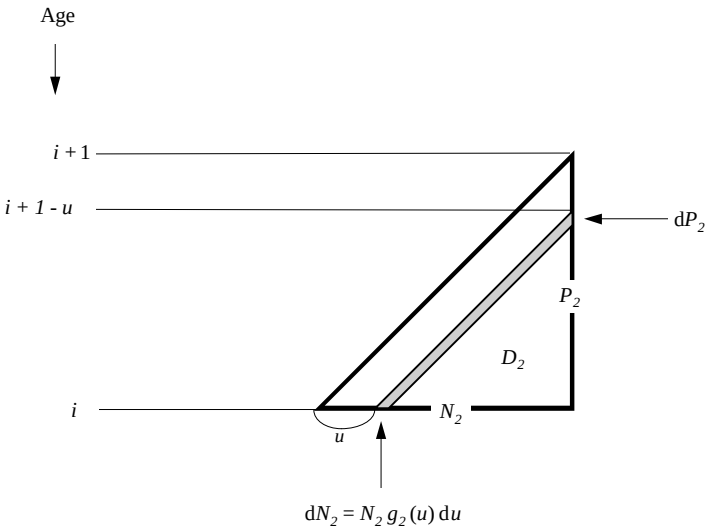


Figure 3. Lower triangle of a square in the Lexis diagram.

Among these $\mathrm{d}N_2$ persons, the number of survivors on December 31st, year T, is :

$$\begin{aligned} N_2\, g_2(u)\, \mathrm{d}u\; \mathrm{e}^{-\int_{x=0}^{1-u} q(x)\,\mathrm{d}x} &= N_2\, g_2(u)\, \mathrm{d}u\; \mathrm{e}^{-\int_{x=0}^{1-u} \left[q + \left(x - \frac{1}{2}\right) q'\right] \mathrm{d}x} \\ &= N_2\, g_2(u)\, \mathrm{d}u\; \mathrm{e}^{-q\,(1-u) - \frac{q'}{2}\left[\left(\frac{1}{2} - u\right)^2 - \frac{1}{4}\right]} \qquad (20) \\ &= N_2\, g_2(u)\, \mathrm{d}u\; \mathrm{e}^{-q\,(1-u) + \frac{q'}{2} u\,(1-u)} \end{aligned}$$

where q and q' refer to exact age i + ½.

On the other hand, $s_2\, \mathrm{d}u\, \mathrm{d}y$ net immigrants, belonging to the same infinitesimal cohort (i.e. the one born in year $T - i$, between times u and $u + \mathrm{d}u$), arrived between ages y and $y + \mathrm{d}y$. Among them, the number of survivors on December 31st of year T, is :

$$\begin{aligned} s_2\, \mathrm{d}u\, \mathrm{d}y\; \mathrm{e}^{-\int_{x=y}^{1-u} q(x)\,\mathrm{d}x} &= s_2\, \mathrm{d}u\, \mathrm{d}y\; \mathrm{e}^{-\int_{x=y}^{1-u} \left[q + \left(x - \frac{1}{2}\right) q'\right] \mathrm{d}x} \\ &= s_2\, \mathrm{d}u\, \mathrm{d}y\; \mathrm{e}^{-q\,(1-u-y) - \frac{q'}{2}\left[\left(\frac{1}{2} - u\right)^2 - \left(y - \frac{1}{2}\right)^2\right]} \qquad (21) \\ &= s_2\, \mathrm{d}u\, \mathrm{d}y\; \mathrm{e}^{-q\,(1-u-y) + \frac{q'}{2}\left[u(1-u) - y(1-y)\right]} \end{aligned}$$

Finally, the number of persons aged i to $i + 1$ on December 31st, is the sum of the integrals corresponding to the two categories :

$$P_2 = N_2 \int_{u=0}^{1} e^{-q(1-u)+\frac{q'}{2}u(1-u)} g_2(u)\,du + s_2 \int_{u=0}^{1}\left[\int_{y=0}^{1-u} e^{-q(1-u-y)+\frac{q'}{2}[u(1-u)-y(1-y)]}\,dy\right]du \quad (22)$$

As shown in Annex 2[53], this equality is approximately :

$$P_2 \approx N_2\ e^{-q(1-m_2)+\frac{q'}{2}m_2(1-m_2)}\left[1+\frac{V_2}{2}(q^2-q')\right] + \frac{s_2}{2}\ e^{-\frac{q}{3}}\left(1+\frac{q^2}{36}\right) \quad (23)$$

where m_2 and V_2 are the *mean* and the *variance*, respectively, of the distribution of the dates of the i^{th} birthdays that are celebrated during calendar year T.
In terms of deaths, (23) can be written, using equation (19), as :

$$\begin{aligned} D_2 &= N_2 - P_2 + \frac{s_2}{2} \\ &\approx N_2\left\{1-e^{-q(1-m_2)+\frac{q'}{2}m_2(1-m_2)}\left[1+\frac{V_2}{2}(q^2-q')\right]\right\} + \frac{s_2}{2}\left[1-e^{-\frac{q}{3}}\left(1+\frac{q^2}{36}\right)\right] \end{aligned} \quad (24)$$

Relationship (24) can be expressed in the following form :

$$q \approx \frac{D_2}{\frac{N_2}{q}\left\{1-e^{-q(1-m_2)+\frac{q'}{2}m_2(1-m_2)}\left[1+\frac{V_2}{2}(q^2-q')\right]\right\} + \frac{s_2}{2q}\left[1-e^{-\frac{q}{3}}\left(1+\frac{q^2}{36}\right)\right]} \quad (25)$$

which is an equation in q for given q', D_2, N_2, s_2, m_2 and V_2. This equation can be solved by applying the rapidly converging iteration relationship :

$$q_{k+1} = \frac{D_2}{\frac{N_2}{q_k}\left\{1-e^{-q_k(1-m_2)+\frac{q'}{2}m_2(1-m_2)}\left[1+\frac{V_2}{2}(q_k^2-q')\right]\right\} + \frac{s_2}{2q_k}\left[1-e^{-\frac{q_k}{3}}\left(1+\frac{q_k^2}{36}\right)\right]} \quad (26)$$

starting from the limit of the right-hand side of (26) when $q \to 0$:

$$q_0 = \frac{D_2}{N_2\left(1-m_2\right)+\frac{s_2}{6}} \quad (27)$$

Thus, based on the assumed prior knowledge of q', D_2, N_2, s_2, m_2 and V_2, we obtain the estimation of risk q corresponding to age $i + ½$. Note that, if q is assumed to be constant ($q' = 0$), the denominator in relationship (26) is equal to the mathematical expectation of the sum of the durations of the periods at risk. The value of q_0, given by (27), is an approximate solution[54] of (24) viewed as an equation in q.

Relationship in the upper triangle

If we consider the *upper* triangle of the square in the Lexis diagram (Figure 4), we have a series of relationships that are similar to relationships (20) to (27).

[53] For a better understanding of Annex 2, Annex 1 should be read prior to Annex 2.
[54] Relationship (27) means that the population at risk approximately consists of two groups :

- N_2 persons, exposed during an *average* period of $1-m_2$ year,
- $\frac{s_2}{2}$ net immigrants, exposed during an *average* period of $\frac{1}{3}$ year.

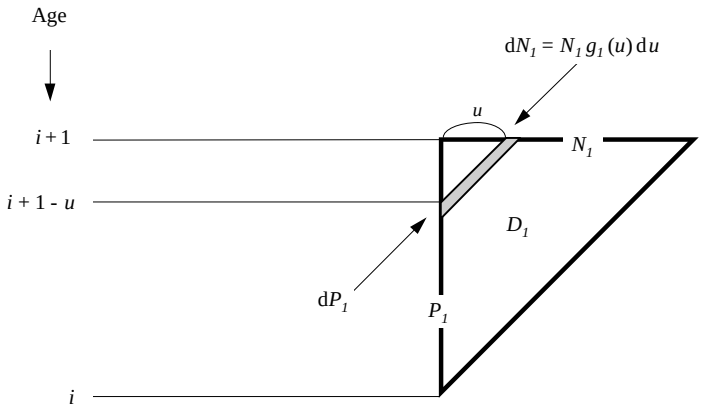


Figure 4. Upper triangle of a square in the Lexis diagram.

Let $dN_1 = N_1\ g_1(u)\ du$ denote the number of persons celebrating their $(i+1)^{th}$ birthday between times u and $u + du$ of year T. Thus $g_1(u)$ is the density of the distribution of $(i+1)^{th}$ birthdays among the surviving and present members of the cohort born during year $T - i - 1$.

These persons belong to *two* categories :

- those already present on the 1st of January, still surviving,
- net immigrants, aged y to $y + dy$ on arrival, who have not died between their immigration and their $(i+1)^{th}$ birthday.

The size of the first subgroup is :

$$dP_1\ e^{-\int_{y=1-u}^{1} q(y)\,dy} = dP_1\ e^{-\int_{x=1-u}^{1}\left[q+\left(x-\frac{1}{2}\right)q'\right]dx} = dP_1\ e^{-qu-\frac{q'}{2}\left[\frac{1}{4}-\left(\frac{1}{2}-u\right)^2\right]} \qquad (28)$$
$$= dP_1\ e^{-qu-\frac{q'}{2}u(1-u)}$$

while that of the second subgroup is :

$$s_1\, du\, dy\ e^{-\int_{y}^{1} q(x)\,dx} = s_1\, du\, dy\ e^{-\int_{y}^{1}\left[q+\left(x-\frac{1}{2}\right)q'\right]dx} = s_1\, du\, dy\ e^{-q(1-y)-\frac{q'}{2}\left[\frac{1}{4}-\left(y-\frac{1}{2}\right)^2\right]} \qquad (29)$$
$$= s_1\, du\, dy\ e^{-q(1-y)-\frac{q'}{2}y(1-y)}$$

Thus we have :

$$N_1\, g_1(u)\, du = dP_1\ e^{-qu-\frac{q'}{2}u(1-u)} + s_1\, du\, dy\ e^{-q(1-y)-\frac{q'}{2}y(1-y)}$$

that is :

$$dP_1 = N_1\ e^{qu+\frac{q'}{2}u(1-u)}\ g_1(u)\, du - s_1\, du\, dy\ e^{qu+\frac{q'}{2}u(1-u)}\ e^{-q(1-y)-\frac{q'}{2}y(1-y)}$$

which leads to :

$$P_1 = N_1 \int_{u=0}^{1} e^{qu+\frac{q'}{2}u(1-u)}\ g_1(u)\, du - s_1 \int_{u=0}^{1} e^{qu+\frac{q'}{2}u(1-u)} \left[\int_{y=1-u}^{1} e^{-q(1-y)-\frac{q'}{2}y(1-y)}\, dy\right] du \qquad (30)$$

As shown in Annex 3, this equality is approximately :

$$P_1 \approx N_1\ e^{qm_1+\frac{q'}{2}m_1(1-m_1)}\left[1+\frac{V_1}{2}\left(q^2-q'\right)\right] - \frac{s_1}{2}\ e^{\frac{q}{3}}\left(1+\frac{q^2}{36}\right) \qquad (31)$$

or, in terms of deaths :

$$D_1 = P_1 - N_1 + \frac{s_1}{2} \approx N_1 \left\{ e^{q\,m_1 + \frac{q'}{2} m_1 (1 - m_1)} \left[1 + \frac{V_1}{2} (q^2 - q') \right] - 1 \right\} - \frac{s_1}{2} \left[e^{\frac{q}{3}} \left(1 + \frac{q^2}{36} \right) - 1 \right] \quad (32)$$

This latter relationship can be written in the following form :

$$q \approx \frac{D_1}{\frac{N_1}{q} \left\{ e^{q\,m_1 + \frac{q'}{2} m_1 (1 - m_1)} \left[1 + \frac{V_1}{2} (q^2 - q') \right] - 1 \right\} - \frac{s_1}{2\,q} \left[e^{\frac{q}{3}} \left(1 + \frac{q^2}{36} \right) - 1 \right]} \quad (33)$$

which is an equation in q for given q', D_1, N_1, s_1, m_1 and V_1. This equation can be solved by applying the rapidly converging iteration relationship :

$$q_{k+1} = \frac{D_1}{\frac{N_1}{q_k} \left\{ e^{q_k m_1 + \frac{q'}{2} m_1 (1 - m_1)} \left[1 + \frac{V_1}{2} (q_k^2 - q') \right] - 1 \right\} - \frac{s_1}{2\,q_k} \left[e^{\frac{q_k}{3}} \left(1 + \frac{q_k^2}{36} \right) - 1 \right]} \quad (34)$$

starting from the limit of the right hand of (34) when $q \to 0$, which constitutes an approximate solution of (32) :

$$q_0 = \frac{D_1}{N_1\ m_1 - \frac{s_1}{6}} \quad (35)$$

Thus, based on the assumed prior knowledge of q', D_1, N_1, s_1, m_1 and V_1, we obtain the estimation of risk q corresponding to age $i + ½$. Again here, if q is assumed to be *constant* ($q' = 0$), the denominator in relationship (34) is equal to the mathematical expectation of the sum of the durations of the periods at risk[55].

Taken together, (24) and (32) constitute the *fundamental relationships* yielding the number of deaths inside each triangle of the Lexis diagram under the assumptions that q' is small compared to q and q^2 is small compared to q'. Except for age 0 – and, to a lesser extent, for age 1 – for which q' is large, such an assumption is close to reality for any age i.

If the relevant birth-cohorts both correspond to a *uniform* density of life lines, we have $m_1 = m_2 = ½$ and $V_1 = V_2 = 1/12$. In this case, (24) and (32) become (36) and (37), respectively :

$$D_2 \approx N_2 \left[1 - e^{\frac{q'}{12}} \frac{1 - e^{-q}}{q} \right] - s_2 \frac{e^{-q} - 1 + q - \frac{q^2}{2}}{q^2} \quad (36)$$

or approximately :

$$q \approx \frac{2\,D_2}{N_2 + \frac{s_2}{3}}$$

and :

[55] Similarly to (27), the expression of q_0 given by (35) means that the population at risk approximately consists of :

- N_1 persons, exposed during an *average* period of m_1 year, *minus*
- $\frac{s_1}{2}$ net immigrants, exposed during an *average* period of $\frac{1}{3}$ year.

$$D_1 \approx N_1 \left[e^{\frac{q'}{12}} \frac{e^q - 1}{q} - 1 \right] - s_1 \frac{e^q - 1 - q - \frac{q^2}{2}}{q^2}. \quad (37)$$

or approximately :

$$q \approx \frac{2 D_1}{N_1 - \frac{s_1}{3}}$$

It is only for birth-cohorts affected by a major event, such as a war (especially the outbreak of a war), that the differences between (24) and (32) on the one hand, and (36) and (37) on the other hand, can be substantial. The values of m_1, m_2, V_1 and V_2 can then be derived from the monthly distribution of births during the *year of birth* of the cohort, under the assumption that mortality and migration have not substantially altered the distribution of birthdays at later ages.

III. – Construction of a period life table (period of one single calendar year)

Let us assume that the following statistical information is available for the population *present* in a given territory, during a given year *T*, by sex and single year of age :

(i) population numbers, on January 1st and December 31st;
(ii) deaths by year of birth, i.e. by *triangle* in the Lexis diagram.

Each square of the Lexis diagram yields *two* estimates of the *force* of mortality q at age $i + ½$: q_1 and q_2 following the iterative sequences (26) and (34). To summarise these two estimates of q, we shall retain the *weighted harmonic mean*, as would be the case if q was assumed to be constant, according to (10). Consequently, the estimate of q in the square is given by the limit of the convergent sequence defined by adding numerators and adding denominators in (26) and (34) :

$$q_{k+1} = \frac{\text{Numerator}}{\text{Denominator}} \quad (38)$$

where :

$$\text{Numerator} = D = D_1 + D_2 = \text{Number of deaths in the square}$$

$$\begin{aligned} \text{Denominator} = {} & \frac{N_1}{q_k} \left\{ e^{q_k m_1 + \frac{q'}{2} m_1 (1 - m_1)} \left[1 + \frac{V_1}{2} (q_k^2 - q') \right] - 1 \right\} \\ & + \frac{N_2}{q_k} \left\{ 1 - e^{-q_k (1 - m_2) + \frac{q'}{2} m_2 (1 - m_2)} \left[1 + \frac{V_2}{2} (q_k^2 - q') \right] \right\} \\ & - \frac{s_1}{2 q_k} \left[e^{\frac{q_k}{3}} \left(1 + \frac{q_k^2}{36} \right) - 1 \right] + \frac{s_2}{2 q_k} \left[1 - e^{-\frac{q_k}{3}} \left(1 + \frac{q_k^2}{36} \right) \right] \end{aligned} \quad (39)$$

An iterative process is necessary to assume that q' is *known* : at the first step, q' is taken as equal to zero ; at the second step, q' is taken as equal to half the difference[56] between the

[56] This estimation of q' corresponds to a parabolic fit over *three* consecutive values of q :

$$q'_i = \frac{q_{i+1} - q_{i-1}}{2}$$

A parabolic fit over *five* consecutive values of q, yielding smoother results, is provided by :

estimates of q yielded by the first step at ages $i + 1$ and $i - 1$, ..., etc. The process is brought to an end when estimates for *all* ages vary little from one step to the next. If it is the *logarithmic* derivative that is assumed to be constant, its value at a given step is taken as half the difference between the logarithms of the estimates of q at ages $i + 1$ and $i - 1$ obtained at the previous step, the first step using zero as an estimate[57].

If the numbers of deaths D are *only* known by *age at last birthday* – i.e. inside each *square*, and not inside each *triangle*, of the Lexis diagram – another iterative process is needed (before the one just mentioned) : at the first step, D_1 and D_2 are taken as equal to $D/2$ and the migration intensities s_1 and s_2 are estimated accordingly. At the second step, D_1 and D_2 are taken as equal to the value derived from the first step by (32) and (24) respectively, and s_1 and s_2 are re-estimated, ..., etc. This process is stopped when D_1 and D_2 are stabilised for *all* ages.

Let us observe that an approximation of the solution of (38) is :

$$\hat{q} \approx \frac{D_1 + D_2}{N_1 m_1 + N_2 (1 - m_2) + \frac{s_2 - s_1}{6}}$$

which shows that the approximate impact of migrations on the number of deaths inside a square of the Lexis diagram is related to the *difference* in the migratory balances inside the two triangles of that square.

Special case of ages 0 and 1

The above procedures apply to any age other than 0 and 1, because of large q' and the impossibility of deriving q' (or q'/q) from the values of q at *both* ages $i - 1$ and $i + 1$. At age 0 for example, we make the three following assumptions :

- there are no migrations,
- during year T, births and first birthdays are *identically* distributed over time within the year,
- D_1 and D_2 are known.

Then Q is estimated by :

$$\hat{Q} = 1 - \left(1 - \frac{D_1}{P_1}\right)\left(1 - \frac{D_2}{N_2}\right) \qquad (40)$$

where $N_2 = B_T$ refers to the number of live births recorded during year T.

Relationship (40) can be proved as follows.

$$q_i' = \frac{2 (q_{i+2} - q_{i-2}) + q_{i+1} - q_{i-1}}{10}$$

[57] The parabolic fit over *five* consecutive values of q yields smoother results :

$$\frac{q_i'}{q_i} = \frac{2 \operatorname{Log}\left(\frac{q_{i+2}}{q_{i-2}}\right) + \operatorname{Log}\left(\frac{q_{i+1}}{q_{i-1}}\right)}{10}$$

Under the three afore-mentioned assumptions, we have :

$$\left(1 - \frac{D_1}{P_1}\right)\left(1 - \frac{D_2}{N_2}\right) = \frac{P_2 / N_2}{P_1 / N_1} = \frac{\int_0^1 e^{-\int_0^{1-u} q(x)\,dx} g_2(u)\,du}{\int_0^1 e^{\int_{1-u}^1 q(x)\,dx} g_1(u)\,du}$$

that is :

$$\frac{\int_0^1 e^{-\int_0^1 q(x)\,dx + \int_{1-u}^1 q(x)\,dx} g_2(u)\,du}{\int_0^1 e^{\int_{1-u}^1 q(x)\,dx} g_1(u)\,du} = (1-Q)\,\frac{\int_0^1 e^{\int_{1-u}^1 q(x)\,dx} g_2(u)\,du}{\int_0^1 e^{\int_{1-u}^1 q(x)dx} g_1(u)\,du}$$

$$= 1 - Q \quad \text{if} \quad \forall u : g_1(u) = g_2(u)$$

This relationship holds for *any* risk function $q(x)$, provided that *migrations* are *zero* and that birthday distributions g_1 and g_2 are *identical* at ages i and $i + 1$. We shall rewrite it as follows:

$$\hat{Q} = 1 - \left(1 - \frac{D_1}{P_1}\right)\left(1 - \frac{D_2}{N_2}\right) \text{ with } a = \frac{D_1}{P_1}, \quad b = \frac{D_2}{N_2}.$$

If net migrations are not zero inside the *first* square but if birthday distributions at ages i and $i + 1$ are identical, let us denote respectively D_1^0 and D_2^0 (instead of D_1 and D_2) the numbers of deaths in the *upper* and the *lower* triangles of the *first* square (i.e. between exact ages 0 and 1), D_1^1 and D_2^1 the similar numbers of deaths in the upper and lower triangles of the *second* square (i.e. between exact ages 1 and 2), D_1^2 and D_2^2 those in the triangles of the *third* square (i.e. between exact ages 2 and 3). Let us also denote respectively P_1^0 and P_2^0 (instead of P_1 and P_2) the population numbers of completed age 0 on January 1st and December 31st, year *T*. Similarly, P_1^1 and P_2^1 are the population numbers of completed age 1 on January 1st and December 31st, year *T*, P_1^2 and P_2^2 those of completed age 2 on January 1st and December 31st, year *T*.

Two estimates of each term *a* and *b* can be calculated. Concerning *a*, i.e. the *upper* triangle, the first estimate assumes that migrations during year *T* within cohort born year *T* - 1 take place *exclusively* on January 1st, the second one that migrations during year *T* within cohort born year *T* take place *exclusively* on December 31st. Under each of these assumptions, net migrations are zero, but in the first one, the population number on January 1st is not P_1^0 but $P_2^1 + D_1^0 + D_2^1$:

$$a_1 = \frac{D_1^0}{P_2^1 + D_1^0 + D_2^1} \qquad a_2 = \frac{D_1^0}{P_1^0}$$

Concerning *b*, i.e. the *lower* triangle, the first estimate assumes that migrations during year *T* within cohort born year *T* - 1 take place *exclusively* immediately after birth, the second one that migrations during year *T* within cohort born year *T* take place *exclusively* on December 31st:

$$b_1 = \frac{D_2^0}{P_2^0 + D_2^0} \qquad b_2 = \frac{D_2^0}{B_T}$$

It follows that the lower and upper bounds of $\hat{Q}$ are:

$1 - [\,1 - \min(a_1, a_2)\,][\,1 - \min(b_1, b_2)\,]$ and $1 - [\,1 - \max(a_1, a_2)\,][\,1 - \max(b_1, b_2)\,]$

If both net migrations during year T respectively inside the vertical-sided parallelogram corresponding to cohort born year T - 1, $s_U = P_2^1 - P_1^0 + D_1^0 + D_2^1$, and inside the lower triangle, $s_L = P_2^0 - B_T + D_2^0$, are small, the two estimates are naturally close to each other. For example, in France in 1990, the lower and upper bounds of the estimated infant mortality rate were respectively 8.41 and 8.48 per 1,000 *male* live births and 6.24 and 6.31 per 1,000 *female* live births.

It must be observed that a satisfactory estimate of the infant mortality rate ($\hat{Q}$ at age 0) requires the availability of the numbers of deaths by *triangle* in the Lexis diagram. If the information available is limited to the *total* number of deaths, SQ_0, inside the *first* square, a rough estimate of the infant mortality rate is still made possible via the preliminary estimation of the breakdown of SQ_0 between D_1^0 and D_2^0. In contemporary Europe, the proportion of D_2^0 to SQ_0 is approximately 90%.

The same procedure can be applied to the estimation of the probability of dying between exact ages 1 and 2. The lower and upper bounds of that probability are again:

$1 - [\,1 - \min(a_1, a_2)\,][\,1 - \min(b_1, b_2)\,]$ and $1 - [\,1 - \max(a_1, a_2)\,][\,1 - \max(b_1, b_2)\,]$

with :

$$a_1 = \frac{D_1^1}{P_2^2 + D_1^1 + D_2^2},\ a_2 = \frac{D_1^1}{P_1^1},\ b_1 = \frac{D_2^1}{P_2^1 + D_2^1},\ b_2 = \frac{D_2^1}{P_1^0 - D_1^0}$$

The lower and upper bounds of the probability of dying between exact ages 1 and 2 are closer (in absolute terms and, frequently, in relative terms) to each other than the lower and upper bounds of the infant mortality rate. For example, in France in 1990, the lower and upper bounds of the probability of dying between exact ages 1 and 2 were respectively 6.97 and 7.02 per 10,000 *male* live births and 5.00 and 5.02 per 10,000 *female* live births.

In contemporary Europe, the share of the lower triangle in the *second* square (between exact ages 1 and 2) is only slightly over 0.5 (around 52 to 55%).

Crude estimate of the probability of dying

If we assume that :

- birthdays are *uniformly* distributed during the year, whatever the birthday considered
- the force of mortality $q(x)$ is *small* and *constant* between exact ages *i* and *i* + 1
- migration intensities, s_1 and s_2, inside the two triangles of the same square, are *equal*
- iterative relationship (38) is applied only *once*, starting with $q \approx 0$
- numbers of deaths D_1 and D_2 are *close* to each other

then *q* is estimated by :

$$\hat{q} \approx \frac{2\,D}{N_1 + N_2} = \frac{2\,D}{P_1 - D_1 + P_2 + D_2} \approx \frac{D}{(P_1 + P_2)/2} \quad (41)$$

and Q by :

$$\hat{Q} = 1 - e^{-\hat{q}} = 1 - e^{-\frac{D}{(P_1 + P_2)/2}} \qquad (42)$$

or even, still *more* crudely :

$$\hat{Q} \approx \frac{D}{(P_1 + P_2)/2} \qquad (43)$$

Note that estimating the force of mortality q by :

$$\hat{q} = \frac{D}{(P_1 + P_2)/2}$$

means estimating an *instantaneous quotient* by the corresponding *observed rate*, i.e. the ratio of the number of events recorded to the *average* exposed population. This *average* exposed population is approximately the *mid-year* population of the age under consideration.

Number of survivors and life expectancy

The estimation $\hat{q}$ of q between ages i and $i + 1$, based on the data available, leads to the following estimate of the probability of dying :

$$\begin{aligned} \hat{Q} &= 1 - e^{-\int_0^1 q(x)\,dx} \\ &= 1 - e^{-\int_0^1 \left[\hat{q} + \left(x - \frac{1}{2}\right) q'\right] dx} \\ &= 1 - e^{-\hat{q}} \end{aligned} \qquad (44)$$

On the other hand, the average time lived between birthdays i and $i + 1$ by survivors at age i is :

$$\begin{aligned} \int_0^1 e^{-\int_0^x q(x)\,dx} dx &= \int_0^1 e^{-\int_0^x \left[\hat{q} + \left(x - \frac{1}{2}\right) q'\right] dx} dx = \int_0^1 e^{-\hat{q}x + \frac{q'}{2} x(1-x)} dx \\ &\approx e^{-\frac{\hat{q}}{2} + \frac{q'}{8}} \left[1 + \frac{\hat{q}^2 - q'}{24}\right] \approx e^{-\frac{\hat{q}}{2} + \frac{q'}{12}} \left[1 + \frac{\hat{q}^2}{24}\right] \\ &\approx e^{\frac{q'}{12}} \frac{1 - e^{-\hat{q}}}{\hat{q}} = e^{\frac{q'}{12}} \frac{\hat{Q}}{\hat{q}} \end{aligned} \qquad (45)$$

If $\hat{Q}_i$ denotes the estimate of the probability of dying at age i, then the number of survivors $S(j \mid i)$ at age j, for 1 survivor at age i, $i < j$, and the life expectancy $e(i)$ at age i are as follows :

$$\begin{aligned} S(j \mid i) &= \prod_{k=i}^{j-1} (1 - \hat{Q}_k) \quad \text{with} \quad S(i \mid i) = 1 \\ e(i) &= \sum_{j >= i} \left[S(j \mid i)\, e^{\frac{q'_j}{12}} \frac{\hat{Q}_j}{\hat{q}_j} \right] \end{aligned} \qquad (46)$$

IV. – Construction of a cohort life table (cohort born during one single year)

If, instead of working on the data for a single calendar year, we use the information concerning the two adjacent triangles of a parallelogram with horizontal sides, for the *same* annual birth-cohort but for *two* consecutive calendar years (Figure 5), then the distribution of birthdays inside the cohort does not play any role.

In this case, the relationship between N_1 and N_2, as can be demonstrated using the results of Annex 1, is :

$$N_1 \approx N_2 \, e^{-q} + \frac{s_1}{2} \, e^{-\frac{q}{3}-\frac{q'}{12}} \left(1+\frac{q^2}{36}\right) + \frac{s_2}{2} \, e^{-\frac{2q}{3}-\frac{q'}{12}} \left(1+\frac{q^2}{36}\right) \qquad (47)$$

or in terms of deaths :

$$\begin{aligned} D &= N_2 - N_1 + \frac{s_1 + s_2}{2} \\ &\approx N_2 \left(1 - e^{-q}\right) + \frac{s_1}{2}\left[1 - e^{-\frac{q}{3}-\frac{q'}{12}}\left(1+\frac{q^2}{36}\right)\right] + \frac{s_2}{2}\left[1 - e^{-\frac{2q}{3}-\frac{q'}{12}}\left(1+\frac{q^2}{36}\right)\right] \end{aligned} \qquad (48)$$

Relationship (48) leads to the converging iterative sequence :

$$q_{k+1} = \frac{D}{\frac{N_2}{2q_k}\left(1 - e^{-q_k}\right) + \frac{s_1}{2q_k}\left[1 - e^{-\frac{q_k}{3}-\frac{q'}{12}}\left(1+\frac{q_k^2}{36}\right)\right] + \frac{s_2}{2q_k}\left[1 - e^{-\frac{2q_k}{3}-\frac{q'}{12}}\left(1+\frac{q_k^2}{36}\right)\right]} \qquad (49)$$

where $D = D_1 + D_2$ is the number of deaths in the parallelogram.

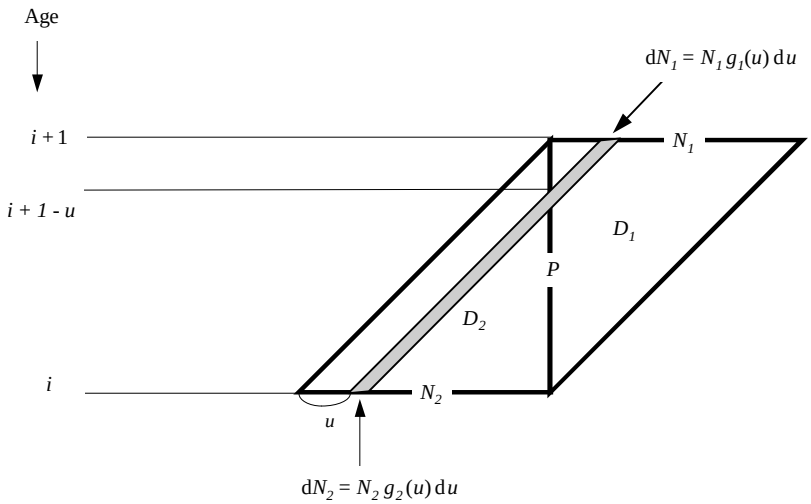


Figure 5. Parallelogram with horizontal sides.

V. – Comparison of methods used to compute life tables

In the case of France, in 1950 and 1990, we have applied the results presented in section III, on the basis of population figures by single year of age at year ends :

- *Method A* : application of iterative relationship (38), knowing the number of deaths by *triangle* in the Lexis diagram and the distribution of birthdays within each annual birth-cohort ;
- *Method B* : application of *double* iterative relationship (38), knowing the number of deaths by *square* in the Lexis diagram and the distributions of birthdays ;
- *Method C* : application of (42), knowing *only* the number of deaths by *square* in the Lexis diagram : the *observed rate* :

$$r = \frac{D}{\left(P_1 + P_2\right)/2}$$

is used as an estimate of the *risk of mortality* and the probability of dying is derived :

$$\hat{q} = r \qquad \hat{Q} = 1 - e^{-r}$$

Two variants of method *C* are considered : methods C_1 and C_2 use the second and the first order approximations, respectively :

$$\hat{Q} = \frac{r}{1 + r/2} \text{ for method } C_1 \qquad \hat{Q} = r \text{ for method } C_2$$

For age 0 and age 1 only, all three methods apply the same procedure : that expressed by relationship (40), assuming that deaths by *triangle* are available for these two special ages.

For the computation of life expectancy at all ages, all four methods assume that, above age 90, probabilities of death increase at a constant rate with age (linear adjustment of their logarithms).

Method A : the recalculation of the number of deaths in each triangle

To assess the validity of Method *A*, which is the most accurate among the three under consideration (if the underlying assumptions hold), we may compare the *observed* numbers of deaths D_1 and D_2 with their *recalculated* values $\hat{D}_1$ and $\hat{D}_2$, derived from $\hat{q}$ according to (32) and (24) :

$$\hat{D}_1 = N_1 \left\{ e^{\hat{q} m_1 + \frac{q'}{2} m_1 (1 - m_1)} \left[1 + \frac{V_1}{2}\left(\hat{q}^2 - q'\right)\right] - 1 \right\} - \frac{S_1}{2}\left[e^{\frac{\hat{q}}{3}}\left(1 + \frac{\hat{q}^2}{36}\right) - 1\right]$$

and :

$$\hat{D}_2 = N_2 \left\{1 - e^{-\hat{q}(1 - m_2) + \frac{q'}{2} m_2 (1 - m_2)} \left[1 + \frac{V_2}{2}\left(\hat{q}^2 - q'\right)\right]\right\} + \frac{S_2}{2}\left[1 - e^{-\frac{\hat{q}}{3}}\left(1 + \frac{\hat{q}^2}{36}\right)\right]$$

Since, by construction, we have :

$$D = D_1 + D_2 = \hat{D}_1 + \hat{D}_2$$

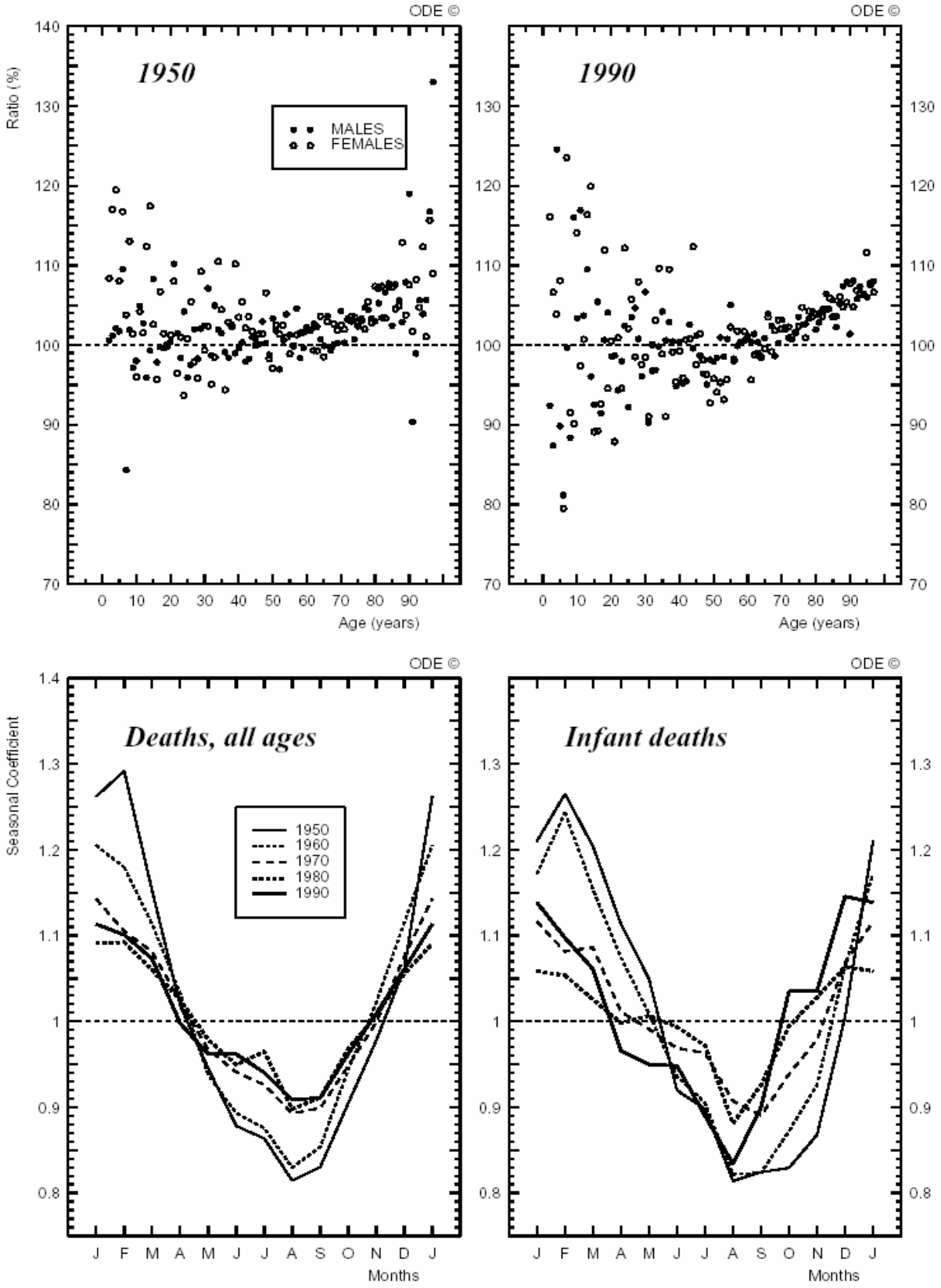
it is sufficient to compare D_1 and $\hat{D}_1$.

Figure 6 shows how the ratio $D_1 / \hat{D}_1$ varies according to age, for males as well as for females, both in 1950 and 1990. It appears that, in 1990, the ratio fluctuates randomly around 1 below age 50. But at higher ages, there is a clear *upward trend* : D_1 is *systematically larger* than it ought to be if the assumptions were fully valid. The same upward trend is also visible in 1950 above age 50, but a *downward trend* seems to exist below 50. In fact, in 1950, the ratio $D_1 / \hat{D}_1$ fluctuates around (slightly *above*) 1 only in the neighbourhood of age 50. The ratio exceeds 1 for all ages more often than not.

The reason for these biases is attributable[58] to the *seasonality* of mortality : the upper triangle contains deaths that mostly occur during the first part of the year while the lower

[58] It might also be due to a wrong allocation, by triangle, of deaths located in a given square of the Lexis diagram.

Figure 6 FRANCE
Upper Part : RATIO, by SEX and AGE, of the OBSERVED NUMBER of DEATHS to the RECALCULATED NUMBER of DEATHS
Method A, UPPER TRIANGLE of each square, 1950 and 1990
Lower Part : SEASONAL PROFILE of the MONTHLY NUMBER of DEATHS
DEATHS for ALL AGES and INFANT DEATHS, 1950, 1960, 1970, 1980, 1990



triangle contains deaths that mostly occur during the latter part of the year. If, at a given exact age, the seasonality of mortality is such that the force of mortality, *q*, is *higher* in the first months than in the last months of the year, then D_1 tends to exceed the value $\hat{D}_1$ estimated under the assumption of absence of seasonality.

If this explanation is true, it means that mortality showed some seasonality – especially above age 50 – at almost all ages in 1950, but this seasonality has greatly declined, if not disappeared, under age 50 in 1990.

For France, deaths by month are available from 1946 to 1991 (for all ages, and for infant deaths), by broad age-groups from 1946 to 1962 and by detailed age-groups from 1981 to 1991. The seasonality of mortality can thus be estimated, for all ages and for age 0 from 1950 to 1991, for age-group 80+ from 1950 to 1958[59] and for detailed age-groups from 1981 to 1991.

Indeed Figure 7 shows that the seasonality of deaths (all ages and age 0) has greatly declined in the past 40 years and that the months with the highest seasonal coefficients are January and February, with November having a seasonal coefficient almost equal to 1. At higher ages, the seasonality is more pronounced than for all ages (January and February had seasonal coefficients for ages 80+ close to 1.5 in the early 1950's). Figure 8 shows the seasonal profile of mortality in selected years. Figure 9 compares the proportion of deaths in the upper triangle of the Lexis diagram to 50 %, using the following index :

$$I = \frac{1}{6} \sum_{m=1}^{12} \left[\left(1 - \frac{2m-1}{24} \right) c_m \right]$$

Index I would be equal to 1 if seasonal coefficients c_m – and death numbers adjusted for seasonal variations – were constant over months. The magnitude of this index and its decline to almost 1 in the 1980's is consistent with the discrepancies shown by Figure 6.

Method B compared to Method A : does the breakdown, by triangle, of the number of deaths in the squares of the Lexis diagram improve the estimation of the probabilities of dying ?

Figure 10 (upper part) shows the percentage difference between the probability of dying estimated according to Method *B* and its equivalent estimated according to Method *A*. In 1950 as well as in 1990, the percentage difference hardly ever reaches 1 %. Below age 70, it is systematically smaller than 0.1 %, and below 85 smaller than 0.5 %. The similar figure for females leads to the same conclusion.

These discrepancies result from the differences between the observed and the recalculated numbers of deaths inside each triangle. It is therefore due to the seasonality of mortality again, but the impact is only noticeable at higher ages.

The conclusion is that the breakdown of the numbers of deaths *D* by triangle adds nothing below 70 years of age – and hardly anything between 70 and 80 years of age – to the quality of the estimation of the probabilities of dying. It is only at ages 80 and over that an appreciable improvement is noted.

However, to establish life tables by birth-cohorts, the breakdown is clearly necessary.

[59] The method used to estimate the seasonal coefficients "loses" four years at each extremity of the period for which monthly data are available. It takes in due consideration the number of days in each month.

Figure 7 FRANCE, 1950-1987
Monthly seasonal coefficients of the number of DEATHS
DEATHS ALL AGES (full lines) and INFANT DEATHS (dotted lines)

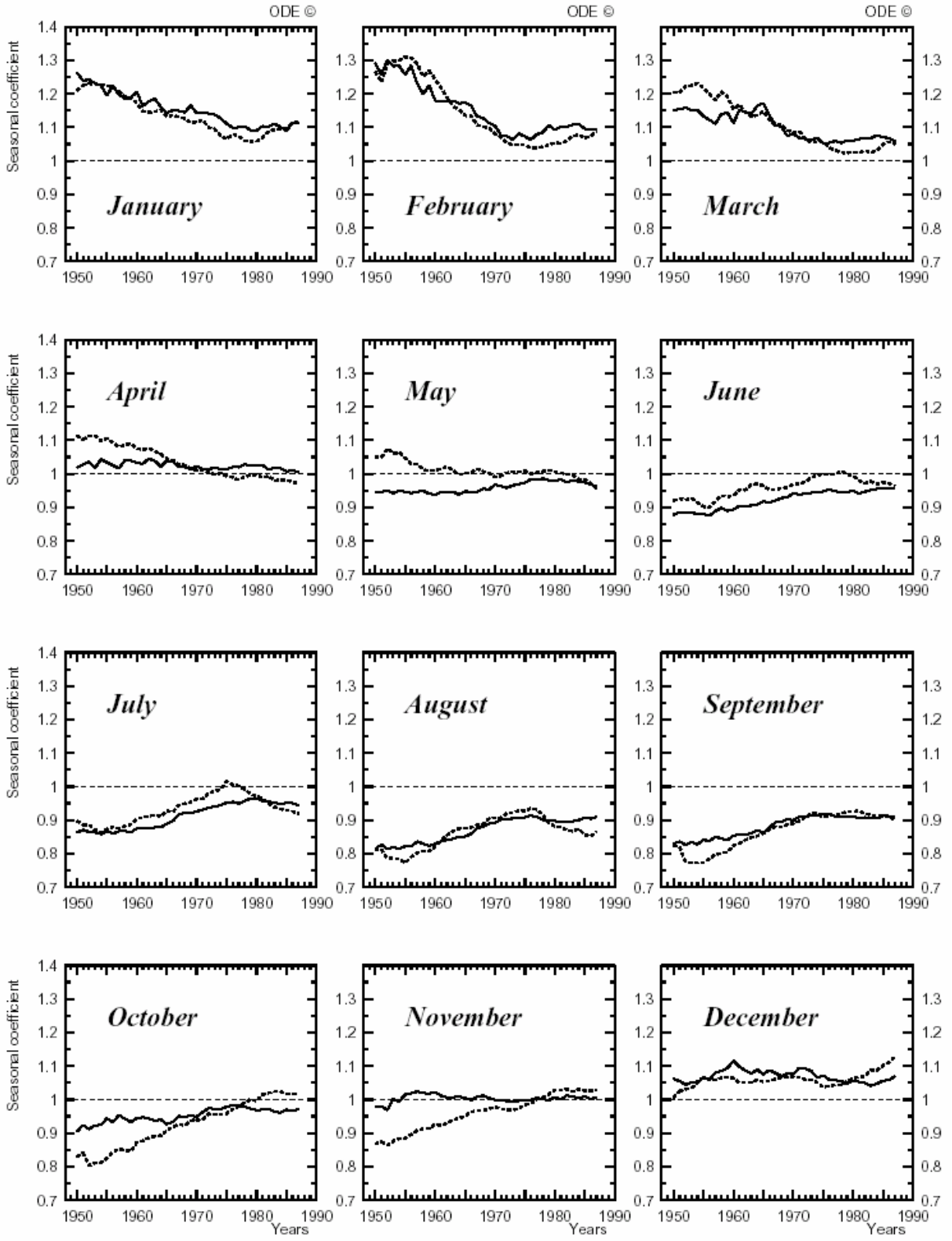


Figure 8 FRANCE, 1950, 1960, 1970, 1980, 1990
SEASONNAL PROFILE of the MONTHLY number of DEATHS
DEATHS for ALL AGES, INFANT DEATHS

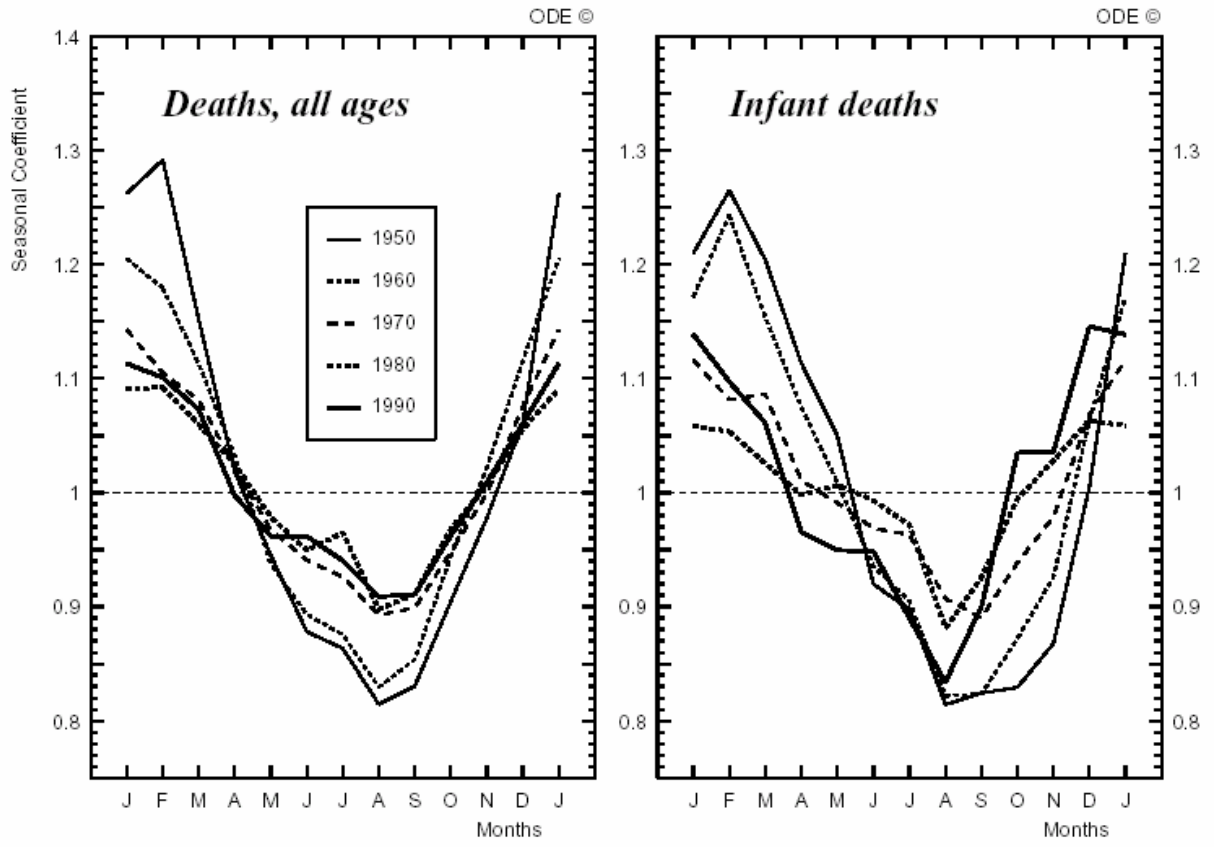


Figure 9 FRANCE, 1950-1987
Ratio of the proportion of DEATHS in the UPPER TRIANGLE to 50 %
(based on seasonal coefficients)

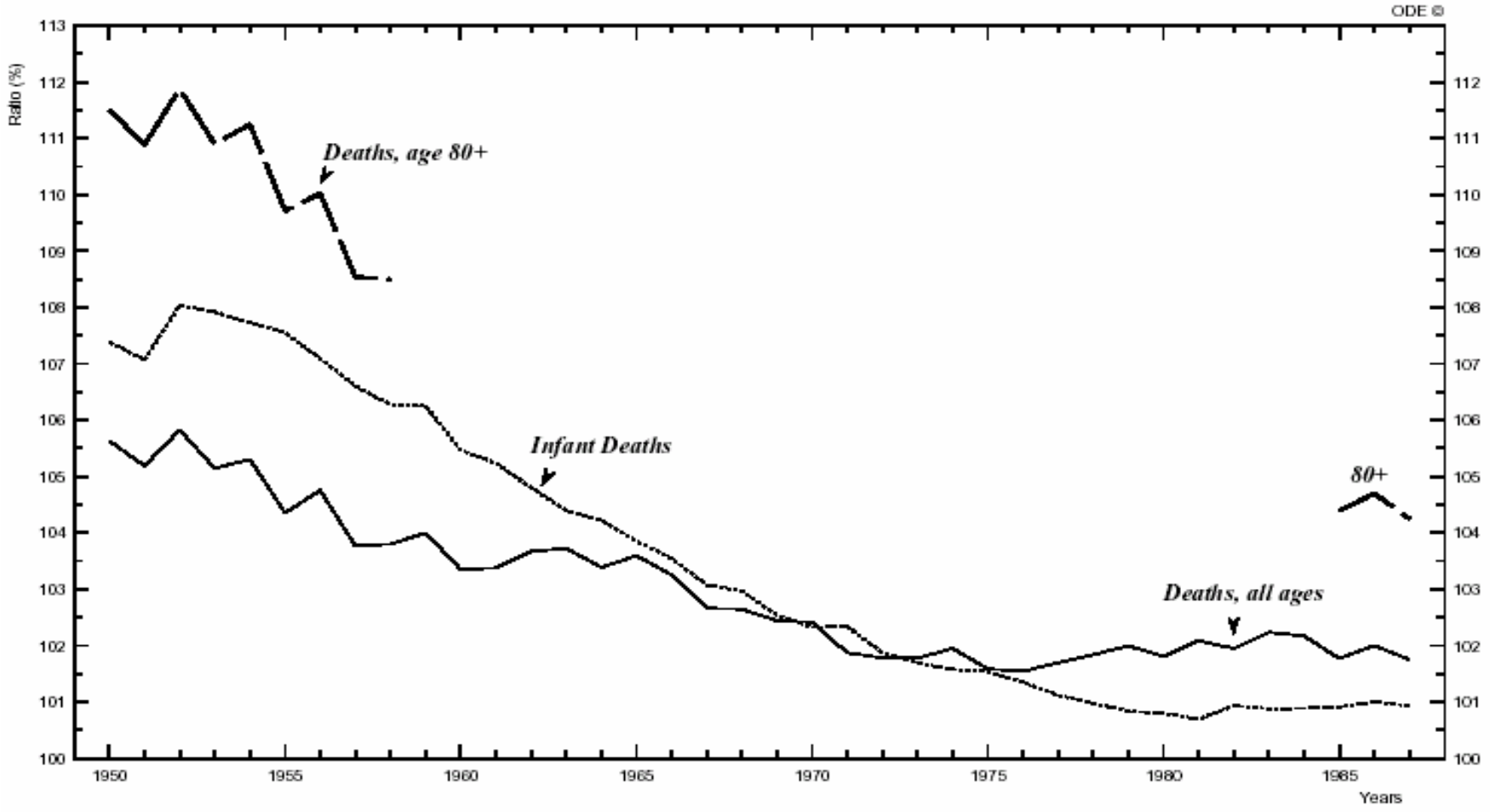
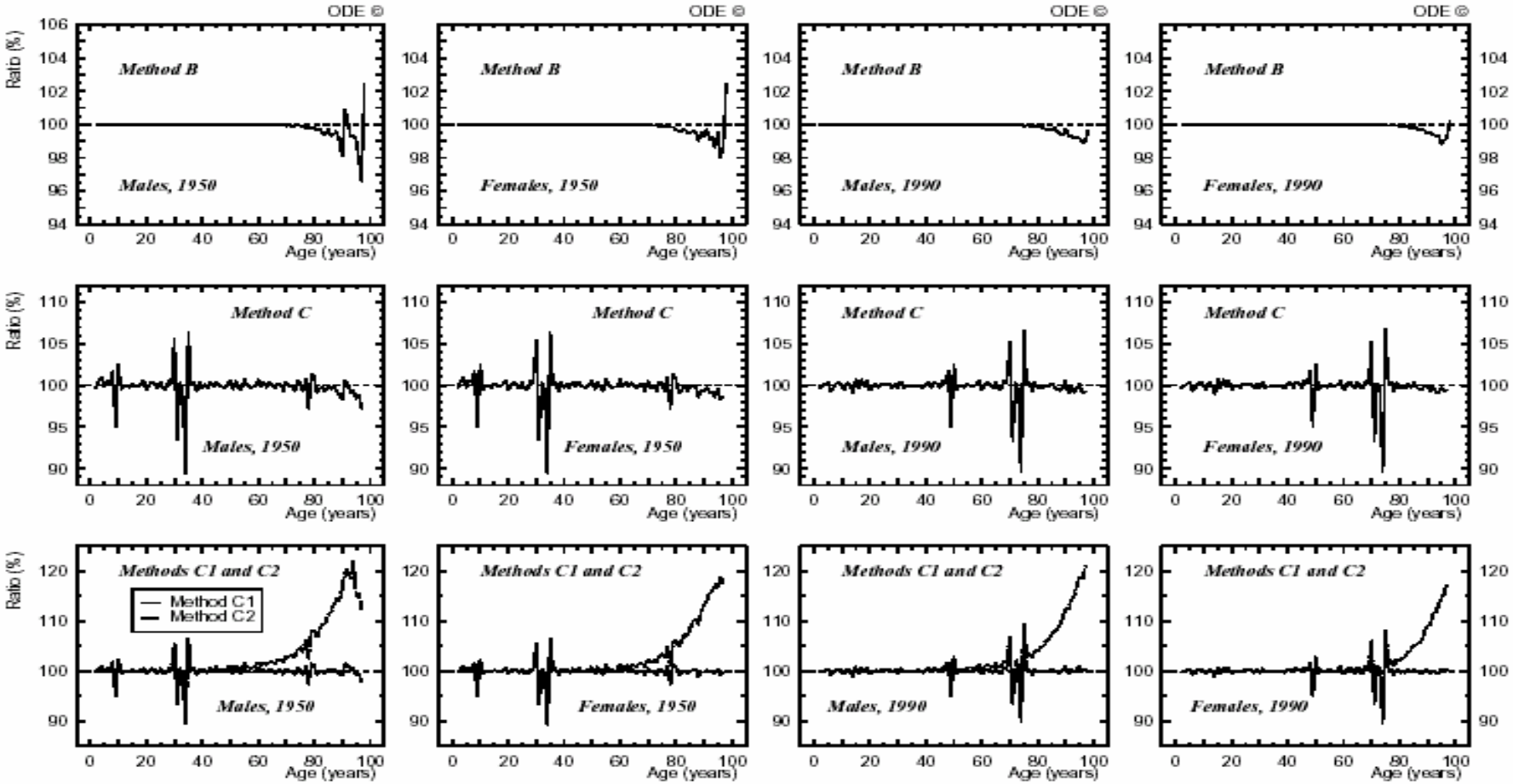


Figure 10 FRANCE, 1950 and 1990
RATIO, by SEX and AGE, of the PROBABILITY of DYING
estimated according to Methods B, C, C1 and C2
to its equivalent estimated according to Method A



Method C and variants C_1 and C_2 compared to Method A : is it worthwhile to take into account the distribution of birthdays within cohorts ?

First, the correction made to remove the effect of the non-uniformity of the distribution of birthdays within the year *can be large*, as is shown in Figures 10 (central and lower parts). For instance, the percentage difference between probabilities of dying according to Method *A* and Method *C* exceeds 10 % at age 34 in 1950 and at age 74 in 1990 (i.e. the 1915-1916 birth-cohorts). With Method C_2, the magnitude of risks (because of the difference between $1-e^{-\phi}$ and ϕ) adds its own effect, bringing the percentage difference with method *A* to 20 % at age 90. But methods *C* and C_1 yield very similar results. The conclusion would be again identical for females.

Second, this correction, which – for a given pair of two consecutive cohorts – is very close to a *constant percent* correction, clearly *attenuates* most of the anomalies which appear in the curves of the probabilities of dying according to age : the correction *smoothes* the curves, as is clear in Figure 11 (upper part).

Third, the magnitude of the correction to apply to the probabilities of dying is large in itself, but also compared to the variations of the latter *over time*. Figure 11 (lower part) shows that several of the erratic fluctuations of *Q* through time are eliminated by the correction.

Finally, it is worthwhile to take the distribution of birthdays into account if the country considered has recorded abrupt changes in its birth-rate, as is the case of most European countries because of the two world wars.

The following question may now be posed : given that the distribution of birthdays is generally only known for the *year of birth*, i.e. for age *zero*, can we assume that mortality and migrations do not substantially modify these distributions for later ages ? This question has been examined in Calot and Caselli (1990)[60], on the basis of census population figures by sex and *month* of birth. The conclusion was that there was good agreement between the corrections based on the distribution of births by month and the corrections based on the distribution of birthdays derived from the census.

VI. – Mortality data for European countries in Eurostat data banks

The analysis of demographic change and, in particular, the assessment of the most *recent* evolution, is a permanent concern for a wide range of observers, at national and international levels. Despite national differences in definitions, *comparable* data are needed – over long periods in the past, but also for most recent periods, even insights into the future through population projections.

This led the then director of the Institut National d'Etudes Démograhiques to set up, in the 1980's, a computerised system for the analysis of demographic data in the field of fertility, which was later extended to nuptiality, and more recently to mortality. This system, called "Projet international d'analyse démographique conjoncturelle", was underpinned with

[60] G. Calot and G. Caselli, Détermination d'une table de mortalité : la conversion des taux en quotients, *Population*, 6, November-December 1991, I.N.E.D., Paris, pp 1441-1490.

information kindly provided by 40 national (or sub-national) statistical offices in Europe and Northern America.

Figure 11 FRANCE, 1990
Upper Part : PROBABILITIES of DYING between 65 and 80 years of age, by SEX and AGE, estimated by Methods A and C
Lower Part : PROBABILITY of DYING at AGE 50 and AGE 60, by SEX
Comparison between Method A and Method C

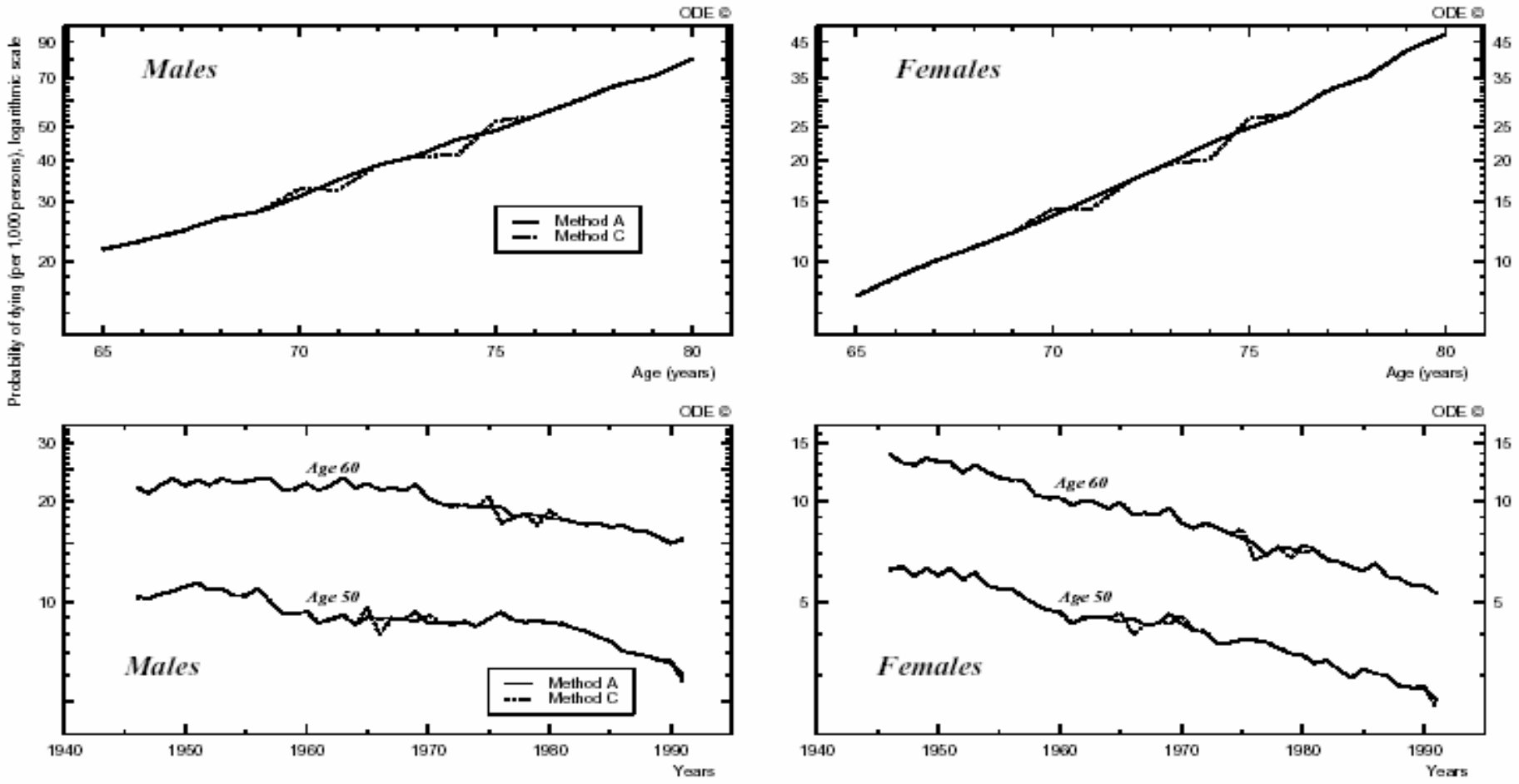


Figure 12 FRANCE, 1950, 1970 et 1990
PROBABILITY of DYING by AGE, according to SEX (Method A)

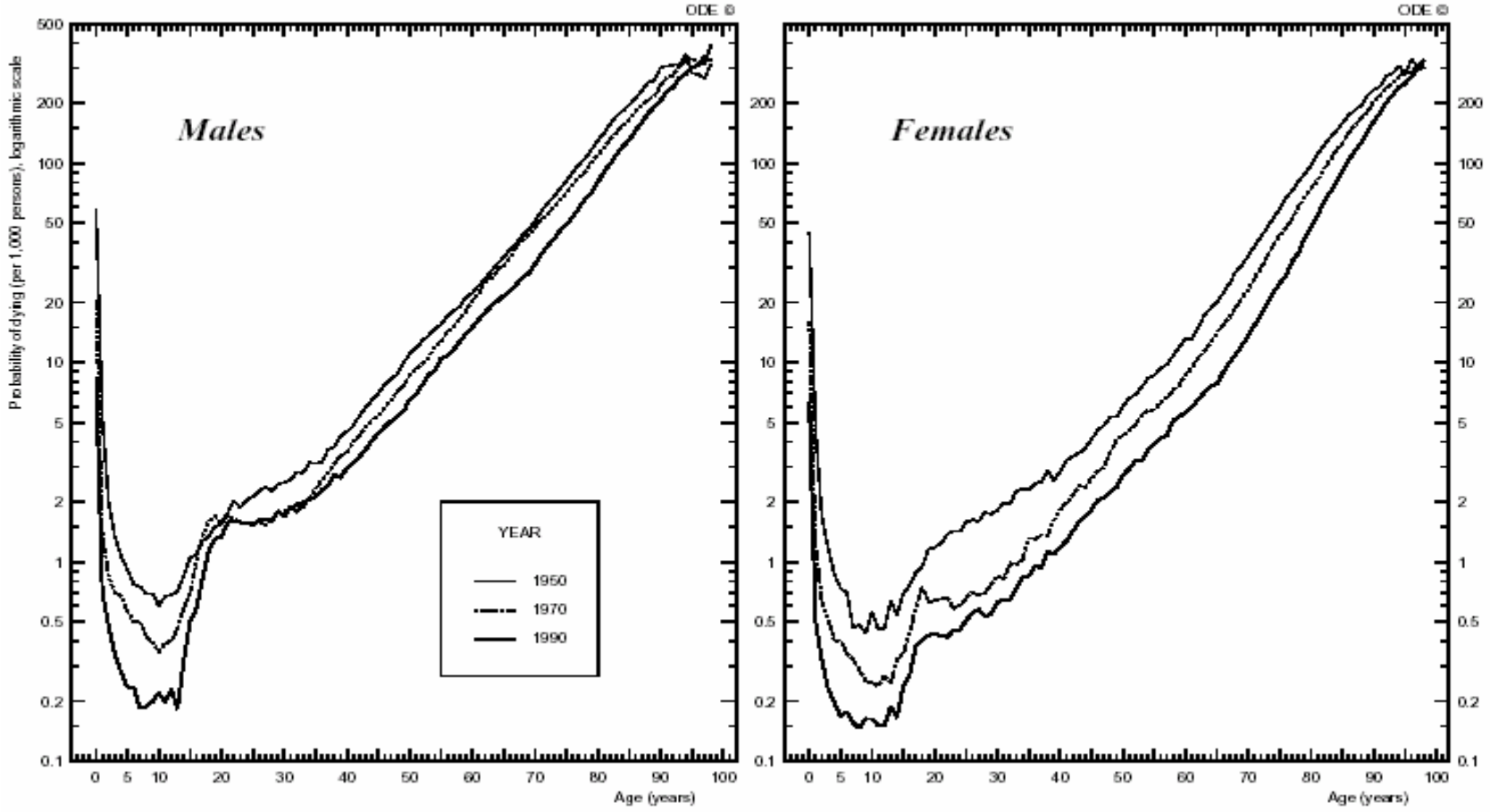
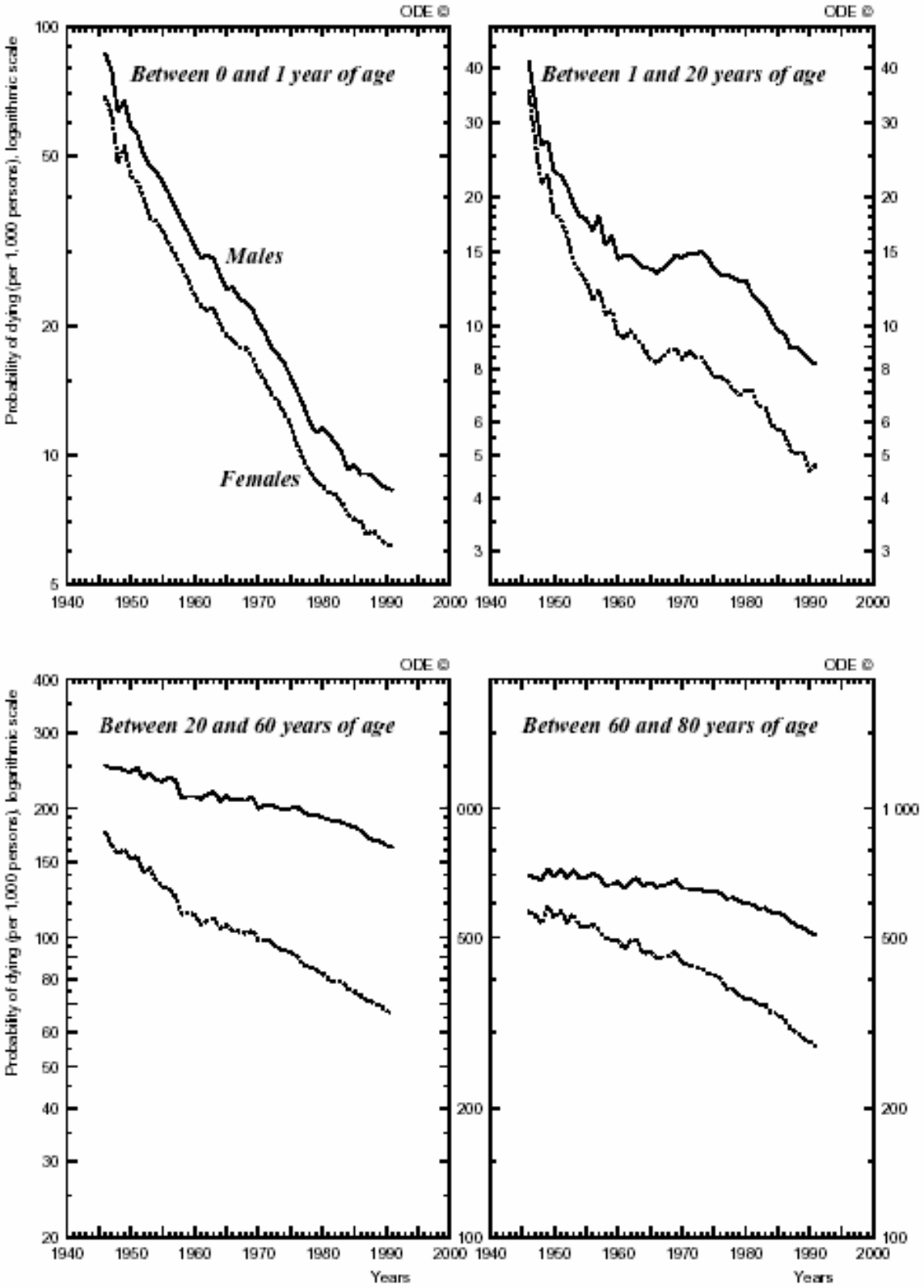


Figure 13 FRANCE, 1946-1991
PROBABILITY of DYING between various ages



The system, together with the data supplied by countries, was transferred to Eurostat, whose responsibility is preparing yearbooks with comparable and up-to-date information on the member States of the European Union, and also keeping the different services of the Commission, the governments and the public at large permanently informed on current changes.

Within Eurostat, the system was renamed *Syscodem* ("Système communautaire d'observation démographique") and the yearly supply of basic information by statistical offices was put on a systematic footing. A software upgrade has since been written. It contains *utility* programs for the acquisition and correction of basic data, *processing* programs which produce comparable outputs, and programs for the *presentation* of the results, in terms of tables and graphs. Examples of the latter[61] are appended to this paper. At the same time, the scope of the system is being extended (in terms of geographical coverage, time periods and fields of study – for example divorces by marriage cohorts).

As far as *mortality by sex and age* is concerned, the basic data supplied by countries for each calendar year are the following :

- *Population by sex and age at year ends*

  The data supplied refer to the resident population on January 1st, by sex and single year of age, up to an open-ended age-group (generally 100+). A few countries produce their data for a different date in the year (for example Ireland on April 15). In this case, estimates are made of the sex-age distributions on January 1st.

- *Deaths by sex, age and year of birth*

  In most countries deaths are classified by age at last birthday *and* year of birth, i.e. by triangle in the Lexis diagram. For a few of them, available data only refer to age at last birthday, i.e. to *squares* in the Lexis diagram.

- *Births by year and month for existing birth-cohorts*
  Live births, not only by year, but also by month, over long periods of the past are available in many countries. These data provide estimates of the distributions of birthdays within cohorts. Census data of total enumerated population by year *and* month of birth can also be used when the monthly distribution of live births is not available.

VII. – Conclusion

The article presents a new approach to the construction of life tables.

It shows that the conventional computation of *rates* based on population numbers at year ends and death numbers by *squares* in the Lexis diagram leads to sizeable *biases* in the probabilities of dying when the birth-rate recorded *abrupt* changes at the time the corresponding cohorts were born[62] (such is the general case in Europe due to the two world wars). On the basis of the *monthly* distributions of live births, commonly available in European countries over long periods of the past, it is possible to correct these biases satisfactorily.

Moreover, the methods proposed *smooth* the probabilities of dying, offering the possibility to establish life tables *annually* (and not only for periods of several years), in spite of the relatively large random fluctuations of the rates due to the small magnitude of death numbers

[61] Figures 6 to 13 were drawn using the Syscodem software for graphs.

[62] These biases are similar to those which occur in the computation of rates for *non-repeatable* events, such as age-specific *fertility* rates.

at most ages. With modern computers, these methods do not require significantly more resources than conventional ones but greatly improve the results.

It is advisable to record annual deaths by sex and single year of age not simply in *squares* of the Lexis triangle, but also in *triangles*. For ages 0 and 1 year of age, it is still more strongly advisable than for other ages.

Annex 1 : Approximations for $\int_a^b e^{j(x)} g(x)\,dx$ *and* $\int_a^b x^k e^{j(x)} g(x)\,dx$

Let us consider, on the one hand, a continuous random variable X, ranging from a to b, with density $g(x)$, mean m, variance V, moments around the mean μ_k and, on the other hand, a continuous function $j(x)$.

The first derivatives of $y = e^{j(x)}$ are :

$$y' = e^{j(x)}\ j'(x)$$
$$y'' = e^{j(x)}\ \left[j'^2(x) + j''(x)\right]$$
$$y''' = e^{j(x)}\ \left[j'^3(x) + 3\,j'(x)\,j''(x) + j'''(x)\right]$$

The development of $e^{j(x)}$ around any m is uniformly convergent. Therefore, denoting j', j'', j''' the successive derivatives of $j(x)$ at $x = m$, the mathematical expectation of $e^{j(X)}$ is :

$$\begin{aligned} E\left[e^{Vj(X)}\right] &= \int_a^b e^{j(x)}\ g(x)\,dx \\ &= \int_a^b e^{j(m)} \left[1 + (x-m)\,j' + \frac{(x-m)^2}{2!}\left(j'^2 + j''\right) + \frac{(x-m)^3}{3!}\left(j'^3 + 3\,j'\,j'' + j'''\right) + \dots\right] g(x)\,dx \\ &\approx e^{j(m)} \left[1 + \frac{V}{2}\left(j'^2 + j''\right) + \frac{\mu_3}{6}\left(j'^3 + 3\,j'\,j'' + j'''\right)\right] \end{aligned}$$

that is :

$$\int_a^b e^{j(x)}\ g(x)\,dx \approx e^{j(m)} \left[1 + \frac{V}{2}\left(j'^2 + j''\right) + \frac{\mu_3}{6}\left(j'^3 + 3\,j'\,j'' + j'''\right)\right] \qquad (1)$$

Similarly, k being a non negative integer, the mathematical expectation of $X^k e^{j(X)}$ is :

$$\begin{aligned} E\left[X^k e^{j(X)}\right] &= \int_a^b x^k\ e^{j(x)}\ g(x)\,dx \\ &= \int_a^b e^{j(m)} \left[m^k + \left(k\,m^{k-1} + m^k\,j'\right)(x-m) + \left(\frac{k(k-1)}{2}\,m^{k-2} + k\,j'\,m^{k-1} + \frac{j'^2 + j''}{2}\,m^k\right)(x-m)^2 + \dots\right] g(x)\,dx \\ &\approx e^{j(m)} \left[m^k + V\,m^{k-2}\left(\frac{k(k-1)}{2} + m\,k\,j' + m^2\,\frac{j'^2 + j''}{2}\right)\right] \end{aligned}$$

that is :

$$\int_a^b x^k\ e^{j(x)}\ g(x)\,dx \approx e^{j(m)} \left[m^k + V\,m^{k-2}\left(\frac{k(k-1)}{2} + m\,k\,j' + m^2\,\frac{j'^2 + j''}{2}\right)\right] \qquad (2)$$

For example, if X is a *uniform* random variable on (0,1), $g(x)$ is equal to 1, m to $\frac{1}{2}$, V to $\frac{1}{12}$ and μ_3 to 0.

We then have :

$$\int_0^1 e^{j(x)}\ dx \approx e^{j\left(\frac{1}{2}\right)} \left(1 + \frac{j'^2 + j''}{24}\right) \qquad (3)$$

and :

$$\int_0^1 x\ e^{j(x)}\,dx \approx \frac{e^{j\left(\frac{1}{2}\right)}}{2} \left(1 + \frac{j'}{6} + \frac{j'^2 + j''}{24}\right) \qquad (4)$$

$$\int_0^1 x^2\ e^{j(x)}\,dx \approx \frac{e^{j\left(\frac{1}{2}\right)}}{3}\left(1 + \frac{j'}{4} + \frac{j'^2 + j''}{32}\right) \tag{5}$$

$$\int_0^1 x^3\ e^{j(x)}\,dx \approx \frac{e^{j\left(\frac{1}{2}\right)}}{4}\left(1 + \frac{j'}{4} + \frac{j'^2 + j''}{48}\right) \tag{6}$$

An illustration of (3) is the following. Let us consider the cumulative function of the *normal* law :

$$\Pi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2p}}\ e^{-\frac{x^2}{2}}\,dx$$
$$= \frac{1}{2} + \frac{u}{\sqrt{2p}}\int_0^1 e^{-\frac{u^2x^2}{2}}\,dx$$

We here have, for any x and therefore for $x = 1/2$:

$$j(x) = -\frac{u^2}{2}x^2 \qquad j = -\frac{u^2}{8}$$
$$j'(x) = -u^2 x \qquad j' = -\frac{u^2}{2}$$
$$j''(x) = -u^2 \qquad j'' = -u^2$$

After (3), we may write :

$$\Pi(u) \approx \frac{1}{2} + \frac{u}{\sqrt{2p}}\ e^{-\frac{u^2}{8}}\left[1 + \frac{u^2}{24}\left(\frac{u^2}{4} - 1\right)\right]$$

This approximation of order 2 for $u = 1$ yields $\Pi(1) = 0.8411$ while the exact value is 0.8413. For $u = 2$, it yields $\Pi(2) = 0.984$ while the exact value is 0.978.

The accuracy of this approximation diminishes as u increases, but improves for fixed u if $e^{j(x)}$ is developed further. The reader may check that the fourth order approximation yields $\Pi(3) = 0.9987$ to four exact decimal places :

$$\Pi(2u) \approx \frac{1}{2} + u\sqrt{\frac{2}{p}}\ e^{-\frac{u^2}{2}} * z(u)$$

with $z(u)$ equal to :

$$1 + \frac{u^2}{3!}\left(u^2 - 1\right) + \frac{u^4}{5!}\left(u^4 - 6u^2 + 3\right) + \frac{u^6}{7!}\left(u^6 - 15u^4 + 45u^2 - 15\right) + \frac{u^8}{9!}\left(u^8 - 280u^6 + 210u^4 - 420u^2 + 105\right)$$

Annex 2 : Approximation for relationship (22)

Let us apply the results of Annex 1 to the two integrals appearing in relationship (22) :

$$P_2 = N_2 \int_{u=0}^{1} e^{-q(1-u)+\frac{q'}{2}u(1-u)}\, g_2(u)\, du + s_2 \int_{u=0}^{1}\left[\int_{y=0}^{1-u} e^{-q(1-u-y)+\frac{q'}{2}\left[u(1-u)-y(1-y)\right]}\, dy\right] du$$

under the assumption that q' and q^2 are small compared to q, q^3 being small compared to q'.

The first integral :

$$N_2 \int_{u=0}^{1} e^{-q(1-u)+\frac{q'}{2}u(1-u)}\, g_2(u)\, du$$

corresponds to :

$$\begin{aligned} j(u) &= -q(1-u) + \frac{q'}{2}u(1-u) \\ j'(u) &= q + q'\left(\frac{1}{2} - u\right) \approx q \\ j''(u) &= -q' \end{aligned}$$

On the basis of relationship (1) in Annex 1, we thus prove the first part of relationship (23) :

$$N_2 \int_{u=0}^{1} e^{-q(1-u)+\frac{q'}{2}u(1-u)}\, g_2(u)\, du \approx N_2\, e^{-q(1-m_2)+\frac{q'}{2}m_2(1-m_2)}\left[1+\frac{V_2}{2}\left(q^2-q'\right)\right]$$

The second integral :

$$s_2 \int_{u=0}^{1}\left[\int_{y=0}^{1-u} e^{-q(1-u-y)+\frac{q'}{2}\left[u(1-u)-y(1-y)\right]}\, dy\right] du$$

can be written :

$$s_2 \int_{u=0}^{1} e^{-q(1-u)+\frac{q'}{2}u(1-u)}\left[\int_{y=0}^{1-u} e^{qy-\frac{q'}{2}y(1-y)}\, dy\right] du$$

To compute the inner integral in y, we apply relationship (3) of Annex 1, putting $y = (1 - u)\, z$:

$$\begin{aligned} \int_{y=0}^{1-u} e^{qy-\frac{q'}{2}y(1-y)}\, dy &= (1-u)\int_0^1 e^{q(1-u)z-\frac{q'(1-u)}{2}z+\frac{q'(1-u)^2}{2}z^2}\, dz \\ &\approx (1-u)\, e^{\frac{q(1-u)}{2}-\frac{q'(1-u)}{4}+\frac{q'(1-u)^2}{8}}\left[1+\frac{\left(q^2+q'\right)(1-u)^2}{24}\right] \\ &= e^{\frac{q(1-u)}{2}-\frac{q'(1-u)}{4}+\frac{q'(1-u)^2}{6}}\left[1-u+\frac{q^2}{24}(1-u)^3\right] \end{aligned}$$

with :

$$\begin{aligned} j(z) &= q(1-u)z - \frac{q'(1-u)}{2}z + \frac{q'(1-u)^2}{2}z^2 \quad \rightarrow \quad j\left(\tfrac{1}{2}\right) = \frac{q(1-u)}{2} - \frac{q'(1-u)}{4} + \frac{q'(1-u)^2}{8} \\ j'(z) &= q(1-u) - \frac{q'(1-u)}{2} + q'(1-u)^2 z \quad \rightarrow \quad j'\left(\tfrac{1}{2}\right) = q(1-u) - \frac{q'(1-u)}{2} + \frac{q'(1-u)^2}{2} \\ j''(z) &= q'(1-u)^2 \end{aligned}$$

It follows that the second term in (22) is equal to :

$$s_2 \int_{u=0}^{1} \left[\int_{y=0}^{1-u} e^{-q(1-u-y)+\frac{q'}{2}[u(1-u)-y(1-y)]}\, dy \right] du = s_2 \int_{u=0}^{1} e^{-q(1-u)+\frac{q'}{2}u(1-u)} \left[\int_{y=0}^{1-u} e^{qy-\frac{q'}{2}y(1-y)}\, dy \right] du$$

$$\approx s_2 \int_{u=0}^{1} e^{-q(1-u)+\frac{q'}{2}u(1-u)}\; e^{\frac{q(1-u)}{2}-\frac{q'(1-u)}{4}+\frac{q'(1-u)^2}{6}} \left[1-u+\frac{q^2}{24}(1-u)^3 \right] du$$

$$= s_2 \int_{u=0}^{1} e^{-\frac{q(1-u)}{2}+\frac{q'(1-u)}{4}-\frac{q'(1-u)^2}{3}} \left[1-u+\frac{q^2}{24}(1-u)^3 \right] du$$

that is :

$$s_2 \int_{u=0}^{1} \left[\int_{y=0}^{1-u} e^{-q(1-u-y)+\frac{q'}{2}[u(1-u)-y(1-y)]}\, dy \right] du \approx s_2 \int_{u=0}^{1} \left(u+\frac{q^2}{24}u^3 \right) e^{-\frac{qu}{2}+q'\left(\frac{u}{4}-\frac{u^2}{3}\right)}\, du$$

To compute this latter expression, we apply relationships (4) and (6) of Annex 1 :

$$s_2 \int_{u=0}^{1} \left(u+\frac{q^2}{24}u^3 \right) e^{-\frac{qu}{2}+q'\left(\frac{u}{4}-\frac{u^2}{3}\right)}\, du \approx \frac{s_2}{2}\, e^{-\frac{q}{4}+\frac{q'}{24}} \left(1-\frac{q}{12}-\frac{q'}{72}+\frac{q^2}{96}-\frac{q'}{36}+\frac{q^2}{48} \right)$$

$$\approx \frac{s_2}{2}\, e^{-\frac{q}{4}} \left(1-\frac{q}{12}+\frac{q^2}{32} \right)$$

$$\approx \frac{s_2}{2}\, e^{-\frac{q}{3}} \left(1+\frac{q^2}{36} \right)$$

with :

$$j(u) = -\frac{qu}{2}+q'\left(\frac{u}{4}-\frac{u^2}{3}\right) \quad \rightarrow \quad j\left(\tfrac{1}{2}\right) = -\frac{q}{4}+\frac{q'}{24}$$

$$j'(u) = -\frac{q}{2}+q'\left(\frac{1}{4}-\frac{2u}{3}\right) \quad \rightarrow \quad j'\left(\tfrac{1}{2}\right) = -\frac{q}{2}-\frac{q'}{12}$$

$$j''(u) = -\frac{2q'}{3}$$

Annex 3 : Approximation for relationship (30)

Let us apply the results of Annex 1 to the computation of the integrals in (30) :

$$P_1 = N_1 \int_{u=0}^{1} e^{q u + \frac{q'}{2} u (1-u)} g_1(u)\, du - s_1 \int_{u=0}^{1} e^{q u + \frac{q'}{2} u (1-u)} \left[\int_{y=1-u}^{1} e^{-q(1-y) - \frac{q'}{2} y (1-y)}\, dy \right] du$$

under the assumption that q' and q^2 are small compared to q, q^3 being small compared to q'.

The first integral :

$$N_1 \int_{u=0}^{1} e^{q u + \frac{q'}{2} u (1-u)} g_1(u)\, du$$

corresponds to :

$$\varphi(u) = q u + \frac{q'}{2} u (1-u)$$

$$\varphi'(u) = q + q' \left(\frac{1}{2} - u \right) \approx q$$

$$\varphi''(u) = -q'$$

Applying relationship (1) of Annex 1, we thus prove the first part of relationship (31) :

$$N_1 \int_{u=0}^{1} e^{q u + \frac{q'}{2} u (1-u)} g_1(u)\, du \approx N_1\, e^{q m_1 + \frac{q'}{2} m_1 (1-m_1)} \left[1 + \frac{V_1}{2} \left(q^2 - q' \right) \right]$$

The second integral is :

$$s_1 \int_{u=0}^{1} e^{q u + \frac{q'}{2} u (1-u)} \left[\int_{y=1-u}^{1} e^{-q(1-y) - \frac{q'}{2} y (1-y)}\, dy \right] du .$$

To compute the integral in y, we apply relationship (3) of Annex 1, putting $y = 1 - u z$:

$$\int_{y=1-u}^{1} e^{-q(1-y) - \frac{q'}{2} y (1-y)}\, dy = u \int_{0}^{1} e^{-q u z - \frac{q'}{2} u z + \frac{q' u^2}{2} z^2}\, dz$$

$$\approx u\, e^{-\frac{q u}{2} - \frac{q' u}{4} + \frac{q' u^2}{8}} \left[1 + \frac{\left(q^2 + q' \right) u^2}{24} \right]$$

$$= e^{-\frac{q u}{2} - \frac{q' u}{4} + \frac{q' u^2}{6}} \left(u + \frac{q^2}{24} u^3 \right)$$

where :

$$\varphi(z) = -q u z - \frac{q'}{2} u z + \frac{q' u^2}{2} z^2 \quad \rightarrow \quad \varphi\left(\tfrac{1}{2}\right) = -\frac{q u}{2} - \frac{q' u}{4} + \frac{q' u^2}{8}$$

$$\varphi'(z) = -q u - \frac{q' u}{2} + q' u^2 z \quad \rightarrow \quad \varphi'\left(\tfrac{1}{2}\right) = -q u - \frac{q' u}{2} + \frac{q' u^2}{2}$$

$$\varphi''(z) = q' u^2$$

It follows that the second term in (30) is equal to :

$$s_1 \int_{u=0}^{1} e^{q u + \frac{q'}{2} u (1-u)} \left[\int_{y=1-u}^{1} e^{-q(1-y) - \frac{q'}{2} y (1-y)} \, dy \right] du \approx s_1 \int_{u=0}^{1} e^{q u + \frac{q'}{2} u (1-u)} \; e^{-\frac{q u}{2} - \frac{q' u}{4} + \frac{q' u^2}{6}} \left(u + \frac{q^2}{24} u^3 \right) du$$

$$= s_1 \int_{u=0}^{1} e^{\frac{q u}{2} + \frac{q' u}{4} - \frac{q' u^2}{3}} \left(u + \frac{q^2}{24} u^3 \right) du$$

$$= s_1 \int_{u=0}^{1} \left(u + \frac{q^2}{24} u^3 \right) e^{\frac{q u}{2} + q' \left(\frac{u}{4} - \frac{u^2}{3} \right)} \, du$$

To compute this latter expression, we apply relationships (4) and (6) of Annex 1 :

$$s_1 \int_{u=0}^{1} \left(u + \frac{q^2}{24} u^3 \right) e^{\frac{q u}{2} + q' \left(\frac{u}{4} - \frac{u^2}{3} \right)} \, du \approx \frac{s_1}{2} \; e^{\frac{q}{4} + \frac{q'}{24}} \left(1 - \frac{q}{12} - \frac{q'}{72} + \frac{q^2}{96} - \frac{q'}{36} + \frac{q^2}{48} \right)$$

$$\approx \frac{s_1}{2} \; e^{\frac{q}{4}} \left(1 - \frac{q}{12} + \frac{q^2}{32} \right)$$

$$\approx \frac{s_1}{2} \; e^{\frac{q}{3}} \left(1 + \frac{q^2}{36} \right)$$

with :

$$\varphi(u) = \frac{q u}{2} + q' \left(\frac{u}{4} - \frac{u^2}{3} \right) \quad \rightarrow \quad \varphi\left(\frac{1}{2}\right) = \frac{q}{4} + \frac{q'}{24}$$

$$\varphi'(u) = \frac{q}{2} + q' \left(\frac{1}{4} - \frac{2 u}{3} \right) \quad \rightarrow \quad \varphi'\left(\frac{1}{2}\right) = \frac{q}{2} - \frac{q'}{12}$$

$$\varphi''(u) = -\frac{2 q'}{3}$$

ANNEX 2

L’analyse démographique conjoncturelle

Reproduit de :Gérard Calot

L’analyse démographique conjoncturelle

in *The joy of demography*,

édité en l’honneur de Dirk J. van de Kaa,

par Anton Kuijsten, Henk de Gans et Henk de Feijter,

NethurD Publications, La Haye, 1999

pp. 295-323

L'analyse démographique conjoncturelle

1. Introduction

La production et la publication des statistiques démographiques courantes, aussi bien en matière de naissances que de mariages ou de décès, est généralement réalisée en plusieurs étapes.

Dans un premier temps, l'Office national de Statistique procède à un comptage des bulletins qu'il reçoit des bureaux locaux d'état civil. A un moment donné, une première estimation *provisoire*, généralement mensuelle, est publiée, qui se fonde sur un nombre de bulletins *déjà* reçus le plus souvent *incomplet* pour diverses sortes de raisons : par exemple parce que les bulletins reçus portent seulement sur une *partie du territoire* (tel est le cas lorsque la périodicité des transmissions de bulletins dépend de la taille de la municipalité) ou parce que le délai maximum légal entre le moment où se produit un événement d'état civil et celui de son enregistrement est de plusieurs mois. Des retards accidentels peuvent également affecter la transmission des bulletins de certains mois.

Dans tous les cas, il est souhaitable que l'Office de Statistique publie non seulement le nombre de bulletins qu'il a effectivement reçus à une date donnée pour un mois *m* donné, mais encore – et surtout – l'estimation (sans biais) qu'il en déduit du nombre *total* d'événements qui se sont produits ce même mois *m* dans l'*ensemble* du pays.

Le cas échéant, une deuxième estimation du nombre d'événements survenus le mois *m*, encore provisoire mais établie sur une base plus exhaustive, donc plus précise, est ultérieurement publiée.

Enfin, avec un décalage chronologique beaucoup plus grand, l'exploitation statistique complète des bulletins d'état civil de l'année conduira à des données *définitives*, en termes non seulement de période d'occurrence des événements mais surtout de *caractéristiques socio-démographiques* des personnes concernées par ces événements (sexe, âge, état matrimonial, ...).

C'est seulement lorsque ces données définitives sont disponibles qu'on peut établir différentes sortes d'indices élaborés permettant d'analyser le phénomène considéré. Un bon nombre de ces indices est obtenu par combinaison des données définitives de *flux* avec les évaluations de population résidente par sexe et âge : tel est le cas des *indicateurs conjoncturels* de *fécondité* ou de *primo-nuptialité* ou de l'*espérance de vie à la naissance*.

Le problème auquel est confronté l'analyste de la conjoncture démographique, dès qu'il dispose d'une estimation, fût-elle provisoire, d'un nombre *absolu*, mensuel ou annuel, d'événements est celui de la *conversion* de cette estimation en celle de l'indice élaboré correspondant. C'est à la solution de ce problème qu'est consacré le présent article.

2. L'effectif moyen des générations soumises au risque

Pour traiter cette question, nous supposerons tout d'abord que l'événement étudié est *renouvelable* et nous prendrons l'exemple des *naissances*. Nous nous appuierons sur le concept d'*effectif moyen des générations soumises au risque* que nous allons en premier lieu présenter.

Considérons, une année donnée, la répartition du nombre des naissances selon l'*âge de la mère*, en admettant que cet âge est défini comme étant celui atteint durant l'année civile de la naissance[63]. Désignons l'année d'observation par *n* (indice supérieur), l'âge de la mère par *i* (indice inférieur), le nombre annuel de naissances de l'année *n* par N^n et le nombre d'enfants nés l'année *n* de mères d'âge *i* (ou, plus précisément, nées elles-mêmes l'année *n* - *i*) par N_i^n (événements situés dans un parallélogramme à côtés verticaux du diagramme de Lexis).

L'effectif des femmes nées l'année *n* – *i* varie légèrement au cours de l'année *n*, sous l'effet de la mortalité et des migrations. Nous conviendrons d'en résumer la valeur pour l'*ensemble* de l'année *n* par la demi-somme de sa valeur au 1er janvier, P_{i-1}^n (effectif d'âge révolu *i* - 1 au 1er janvier *n*), et de celle, P_i^{n+1}, au 31 décembre :

$$F_i^n = \frac{P_{i-1}^n + P_i^{n+1}}{2}$$

Le *taux de fécondité* f_i^n à l'âge *i* atteint pendant l'année *n* est le *rapport* entre le nombre de naissances N_i^n et l'effectif correspondant soumis au risque F_i^n :

$$f_i^n = \frac{N_i^n}{F_i^n} = \frac{N_i^n}{\left(P_{i-1}^n + P_i^{n+1}\right)/2}$$

L'*indicateur conjoncturel* de fécondité I^n de l'année *n* est la *somme*, étendue aux différents âges de la période féconde (15-49 ans, pour fixer les idées), des taux de fécondité par âge :

$$I^n = \sum_{i=15}^{49} f_i^n$$

Il s'ensuit que le rapport G^n du nombre annuel de naissances N^n à l'indicateur conjoncturel I^n est la moyenne *pondérée* des effectifs féminins F_i^n aux divers âges de fécondité, le coefficient de pondération de l'effectif d'âge *i* étant le taux de fécondité f_i^n à ce même âge *i* observé cette même année *n* :

$$G^n = \frac{N^n}{I^n} = \frac{\sum_{i=15}^{49} N_i^n}{\sum_{i=15}^{49} f_i^n} = \frac{\sum_{i=15}^{49} f_i^n F_i^n}{\sum_{i=15}^{49} f_i^n}$$

soit encore :

$$N^n = I^n . G^n$$

[63] C'est seulement pour faciliter l'exposé que nous supposons que l'âge de la mère est celui *atteint durant l'année civile de la naissance*. Si l'âge de la mère est celui *en années révolues au moment de la naissance*, on aboutit à des résultats équivalents : le taux de fécondité f_i^n à l'âge *i* pour l'année *n* est alors le rapport entre le nombre N_i^n d'événements observés dans le *carré* du diagramme de Lexis et le nombre de femmes-années d'exposition au risque F_i^n, estimé par $\left(P_i^n + P_i^{n+1}\right)/2$ en admettant que la densité des lignes de vie au sein des deux générations annuelles concernées est *uniforme* :

$$f_i^n = \frac{N_i^n}{\left(P_i^n + P_i^{n+1}\right)/2}$$

Le nombre annuel N^n de naissances de l'année *n* apparaît ainsi comme le *produit* de l'indicateur conjoncturel I^n, résumé des comportements de fécondité *propres* à l'année *n*, par l'effectif moyen pondéré G^n des générations féminines qui, au cours de l'année *n*, appartiennent aux âges féconds. Cet effectif moyen G^n est l'*héritage* de la natalité des années de 15 à 49 ans antérieures à l'année *n* considérée, corrigé par la mortalité et les migrations intervenues depuis l'époque de la naissance.

En d'autres termes, le nombre de naissances de l'année *n* est la résultante *multiplicative* d'une *intensité*, caractéristique des comportements de fécondité de l'année *n* elle-même, intensité que mesure l'indicateur conjoncturel, et d'un *effectif*, hérité du passé, égal à la moyenne *pondérée* des effectifs féminins F_i^n de l'année considérée selon l'âge, le poids de l'effectif d'âge *i* étant le taux de fécondité f_i^n correspondant à cet âge et à cette année.

En toute rigueur, l'effectif moyen G^n, que nous qualifions d'*héritage* du passé, dépend un peu des comportements de fécondité de l'année *n* puisque les coefficients de pondération des effectifs F_i^n sont les taux de fécondité f_i^n de l'année *n elle-même*. Mais on sait que la moyenne pondérée d'éléments pas trop variables (les effectifs F_i^n) dépend assez peu des coefficients de pondération (les taux f_i^n) : si on modifie légèrement ceux-ci, on ne modifie guère la valeur de la moyenne pondérée. En particulier, lorsque les effectifs F_i^n de l'année *n* ne dépendent pas de l'âge *i*, c'est-à-dire lorsque les effectifs féminins aux divers âges de fécondité sont de *même taille*, l'effectif moyen pondéré coïncide avec la valeur commune des F_i^n, *quels que soient* les coefficients de pondération. Aussi ne commet-on pas grande erreur en raisonnant comme si l'effectif moyen G^n dépendait *exclusivement* des effectifs F_i^n et *aucunement* des taux f_i^n.

On aura une idée de l'effet des changements de coefficients de pondération en considérant la figure 1 qui décrit, pour la France, l'évolution depuis 1946 de l'effectif moyen des générations féminines en âge de fécondité, établi respectivement en utilisant pour coefficients de pondération les taux de l'année *n* elle-même (valeur exacte) et les taux d'une année *fixe* (successivement : 1950, 1960, 1970, 1980 et 1990). C'est dans la période 1965-1985, époque où sont parvenues aux âges de fécondité les générations du *baby-boom* et où, par conséquent, les effectifs des divers âges féconds étaient le plus *inégaux*, que l'effet est maximum. L'effet est sensiblement nul vers 1990 parce que les générations féminines d'âge fécond sont, à cette date, de taille très voisine.

Cas particulier :
mortalité et migrations nulles avant 50 ans et calendrier transversal de fécondité invariable

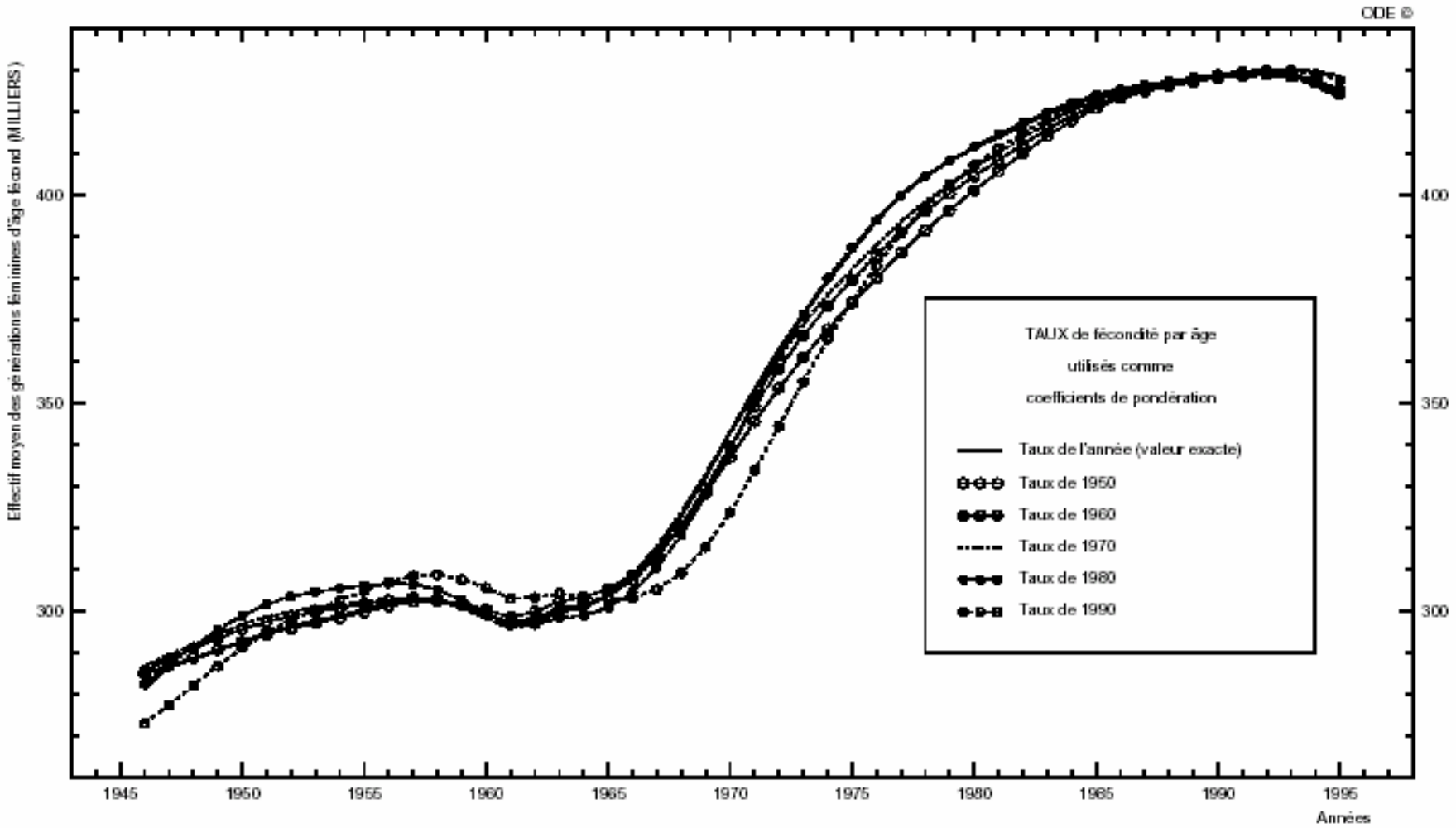
Supposons que la mortalité et les migrations sont *nulles* entre la naissance et la fin de la période féconde. Alors, le nombre F_i^n de femmes d'âge *i* l'année *n* est égal au nombre de naissances *féminines* de l'année *n* – *i*, soit encore $\varphi\ N^{n-i}$, φ étant la proportion de filles parmi les nouveau-nés.

Supposons, de plus, que le *calendrier* de la fécondité transversale est *invariable* d'une année à l'autre :

$\frac{f_i^n}{I^n}$ dépend de *i* mais non de *n*, soit $\frac{f_i^n}{I^n} = a_i$, avec $\sum_{i=15}^{49} a_i = 1$

et désignons par $\bar{x}$ l'*âge moyen* constant à la maternité auquel correspond ce calendrier invariable :

Figure 1 FRANCE, 1946-1995
Effectif moyen des générations féminines d'âge fécond
Valeur exacte et valeurs estimées sur la base des taux d'une année donnée



$$\bar{x} = \sum_{i=15}^{49} \alpha_i \, i$$

Sous ces deux hypothèses, la série G^n de l'effectif moyen des générations féminines d'âge fécond est, au coefficient φ près, la *moyenne mobile pondérée* de la série des naissances annuelles, *décalée* de l'âge moyen constant à la maternité $\bar{x}$:

$$G^n = \varphi \sum_{i=15}^{49} \alpha_i N^{n-i} = \varphi \sum_{k=n-49}^{n-15} \alpha_{n-k} N^k$$

$$n - \bar{x} = \sum_{i=15}^{49} \alpha_i (n-i) = \sum_{k=n-49}^{n-15} \alpha_{n-k} \, k$$

Ces deux hypothèses simplificatrices ne sont jamais très éloignées de la réalité, ce qui explique que l'évolution de la courbe G^n est généralement très *lisse*, comme l'est celle d'une moyenne mobile sur un assez grand nombre de termes (on a ici 35 termes : de 15 à 49 ans). Ce caractère lisse de l'évolution de G^n va faciliter les interpolations et les extrapolations[64].

Comme on le constate sur la figure 2, qui décrit l'évolution observée depuis la Seconde guerre mondiale dans six pays européens, les variations de l'effectif moyen G^n des générations féminines d'âge fécond sont effectivement régulières.

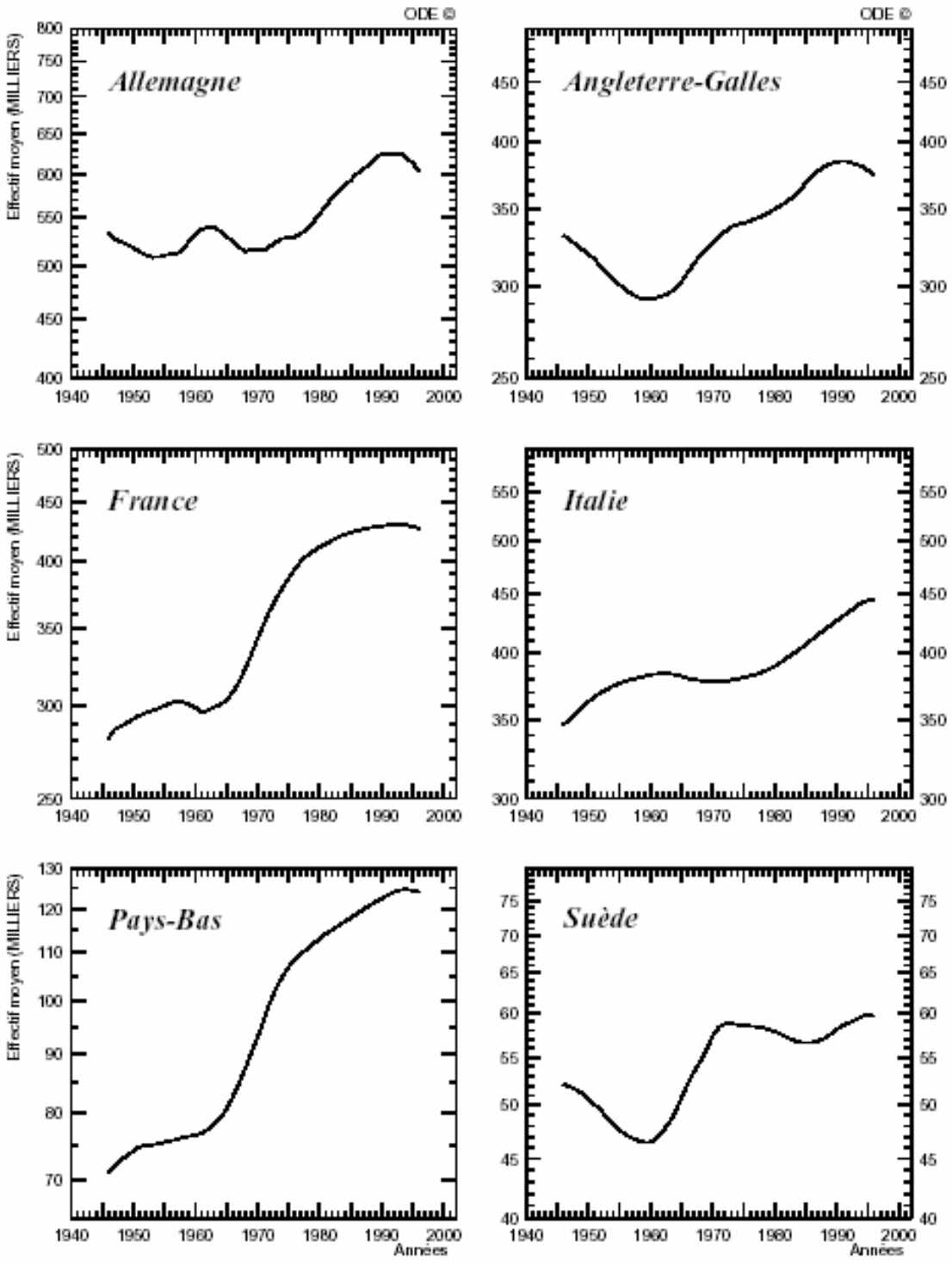
On notera l'effet, variable selon les pays – en ampleur et, dans une moindre mesure, en calendrier –, qu'a eu le *baby-boom* sur l'augmentation de cet effectif moyen. Par ailleurs, dans la plupart des pays considérés, l'effectif moyen a commencé à décroître vers 1990, en écho à la chute du nombre absolu des naissances à partir des années 1965-1970 : si, depuis vingt-cinq ans, le nombre absolu des naissances en Europe a décru moins rapidement, en valeur relative, que l'indicateur conjoncturel de fécondité, dans les années futures, la diminution de l'effectif moyen des générations féminines d'âge fécond pèsera, à la baisse, sur l'évolution du nombre absolu des naissances.

Définition générale de l'effectif moyen des générations soumises au risque

D'une façon générale, considérons un flux annuel N^n d'événements renouvelables, classés selon l'âge *i* de la personne qui le subit, et définissons le taux f_i^n à l'âge *i* comme le rapport du nombre N_i^n d'événements au nombre de personnes-années F_i^n d'exposition au risque à l'âge *i* durant l'année *n* au sein de la population *totale*. Alors, le flux annuel N^n est le produit de la somme des taux f_i^n par la moyenne pondérée des nombres F_i^n, c'est-à-dire par l'*effectif moyen* G^n *des générations soumises au risque*. Observons que si l'événement considéré n'est pas renouvelable (ainsi, le *premier* mariage) mais traité comme renouvelable, les taux f_i^n sont les taux dits de *seconde catégorie*, qui utilisent comme dénominateur l'effectif *total* de la population résidente d'âge *i* et non l'effectif des *seules* personnes (dans l'exemple du premier mariage : les *célibataires*) qui n'ont pas enregistré l'événement et qui sont pourtant les seules à être effectivement soumises au risque.

[64] En matière de *projection* de population, on peut s'appuyer sur le caractère lisse de l'évolution du rapport entre le nombre absolu projeté de naissances et l'indicateur conjoncturel de fécondité également projeté, pour détecter d'éventuelles erreurs de calcul.

Figure 2
EFFECTIF MOYEN des GENERATIONS FEMININES
en AGE de FECONDITE dans six pays européens
Echelles des ordonnées logarithmiques de même module



3. L'effectif moyen des générations en âge de primo-nuptialité

En matière de primo-nuptialité, c'est-à-dire de mariages de célibataires, on peut déterminer, de façon analogue au cas de la fécondité, l'effectif moyen des générations *masculines* et l'effectif moyen des générations *féminines* en âge de primo-nuptialité, en traitant le *premier* mariage comme un événement renouvelable.

Ces deux effectifs moyens n'évoluent pas de manière rigoureusement simultanée (figure 3), bien que les variations des effectifs de population masculine et féminine soient généralement concomitantes et de même ampleur : la raison tient au fait que l'âge moyen au premier mariage des femmes est plus *précoce* (de l'ordre de deux à trois années) que celui des hommes. Il en résulte qu'une augmentation de natalité provoque, une vingtaine d'années après, une augmentation de l'effectif féminin moyen en âge de primo-nuptialité deux à trois ans *plus tôt* que celle de l'effectif masculin. La présence temporaire d'un plus grand nombre de candidates au premier mariage que de candidats a pour effet d'abaisser l'indicateur conjoncturel de primo-nuptialité féminine et de relever son homologue masculin (figure 4), la situation inverse produisant un effet inverse.

Sur la figure 4, on notera le cas particulier de l'Allemagne, où l'augmentation de la natalité de la période 1934-1944 a provoqué, vers 1960, une augmentation de l'effectif féminin en âge de premier mariage quelques années plus tôt que celle de l'effectif masculin. En outre, dans ce pays, les pertes militaires de la Seconde guerre mondiale sont à l'origine de l'excédent considérable de candidates au premier mariage de 1945 à 1955.

Les variations temporelles de l'effectif moyen des générations en âge de primo-nuptialité font ainsi apparaître les « tensions » qui se manifestent sur le « marché matrimonial » du fait des évolutions non rigoureusement parallèles des effectifs de l'un et l'autre sexe.

4. L'interpolation mensuelle de l'effectif moyen annuel

L'interpolation, à l'échelle mensuelle, de l'effectif moyen annuel des générations soumises au risque est facilitée par le caractère *lisse* de ce dernier. En convenant que le douzième de la valeur annuelle est la valeur mensuelle *typique* de l'année, qui se situe à mi-chemin entre juin et juillet, on peut définir une courbe *régulière* passant par ces valeurs typiques, puis lire les valeurs de chaque mois sur la courbe régulière ainsi déterminée.

C'est ce qui a été réalisé sur la figure 5 qui se rapporte à l'effectif moyen des générations féminines de la France en âge de fécondité. On a ajusté une courbe polynômiale de degré 5 sur six points typiques consécutifs (juin-juillet des années n+1 à n+6) et, pour les douze mois de la période *centrale* (qui va de juillet n+3 à juin n+4), on a retenu les valeurs mensuelles lues sur cette courbe polynômiale ajustée[65].

Grâce au degré élevé des polynômes utilisés, la courbe mensuelle ajustée passe exactement par les valeurs annuelles typiques observées et les raccords d'une période centrale à la suivante se font *sans* discontinuité, en termes aussi bien de valeurs que de dérivées d'ordre 1 ou 2. C'est seulement à chacune des *extrémités* de la période d'étude qu'on retient *aussi*, au moins *provisoirement*, les valeurs mensuelles lues sur la courbe

[65] De façon précise, considérons six années consécutives, soit une période de 72 mois. Prenons pour date-origine le 1er janvier de la troisième année et adoptons le mois comme unité de durée. Les valeurs typiques des six années se rapportent aux dates -30, -18, -6, 6, 18, 30 et les milieux des mois de la période centrale se situent aux dates -5,5 (juillet de la troisième année), -4,5, ..., 4,5, 5,5 (juin de la quatrième année). L'ajustement polynômial consiste à déterminer la courbe de degré 5 qui passe par les six points d'abscisses ±6, ±18, ±30 et à retenir les douze valeurs correspondant aux abscisses ±0,5, ±1,5, ..., ±5,5.

polynômiale ajustée, pour les mois respectivement *antérieurs* à la *première* période centrale et *postérieurs* à la *dernière* période centrale. Lorsqu'on disposera ultérieurement d'informations supplémentaires, on *révisera* les valeurs correspondant aux nouvelles périodes centrales. Par ailleurs, de façon à améliorer la qualité de l'ajustement pour les mois du passé récent ou du futur proche, on peut procéder, préalablement à l'interpolation mensuelle, à une *extrapolation* des valeurs annuelles.

5. L'extrapolation de l'effectif moyen annuel

A un moment donné, désignons par *a* l'année *la plus récente* pour laquelle les taux de fécondité par âge f_i^n sont disponibles et par *b* (avec généralement *b* >= *a*+1) l'année *la plus récente* pour laquelle on connaît les effectifs féminins au 1er janvier P_i^n par âge révolu.

Pour les années *a* et avant, on connaît le nombre absolu d'événements N^n, la somme I^n des taux f_i^n et donc le rapport $G^n = N^n / I^n$. Comment estimer l'effectif moyen G^n pour les années *n* postérieures[66] à *a*, en supposant que le décalage de *n-a* années n'est pas trop grand (disons *n-a* au plus égal à 5 ou 10 ans) ?

Une première méthode, purement graphique, consiste en une extrapolation *manuelle* à l'année *n* de la courbe *lisse* G^k connue jusqu'à l'année *k = a*.

On peut aussi procéder par le calcul et convenir de remplacer, dans l'expression de G^n :

$$G^n = \frac{\sum_{i=15}^{49} f_i^n F_i^n}{\sum_{i=15}^{49} f_i^n} = \frac{\sum_{i=15}^{49} f_i^n \frac{P_{i-1}^n + P_i^{n+1}}{2}}{\sum_{i=15}^{49} f_i^n}$$

le taux non encore observé f_i^n par le *dernier* observé *au même âge* (en notant que ce taux est *obsolète* de *n-a* années), soit f_i^a, et l'effectif non encore observé F_i^n par le *dernier* observé *pour la même génération*, soit $P_{i-1-(n-b)}^b$ (cet effectif est *obsolète* de *n-b+½* années) :

$$\hat{G}^n = \frac{\sum_{i=15}^{49} f_i^a P_{i-1-(n-b)}^b}{\sum_{i=15}^{49} f_i^a}$$

Le rapport de la valeur *estimée*, résultant de l'application de ce type de formule, à la valeur *exacte* peut être mesuré pour une année *k* quelconque, *antérieure* ou égale à *n-a*, sur la base de taux obsolètes de *n-a* années et d'effectifs obsolètes de *n-b+½* années :

$$\frac{\hat{G}^k}{G^k} = \frac{\sum_{i=15}^{49} f_i^{k-(n-a)} P_{i-1-(n-b)}^{k-(n-b)}}{\sum_{i=15}^{49} f_i^{k-(n-a)}} \Big/ \frac{\sum_{i=15}^{49} f_i^k \frac{P_{i-1}^k + P_i^{k+1}}{2}}{\sum_{i=15}^{49} f_i^k}$$

[66] Nous traitons ici de l'extrapolation (vers le *futur*) de l'effectif moyen G^n. Le problème serait très voisin si on voulait rétropoler (vers le *passé*) cet effectif moyen.

Figure 3
EFFECTIF MOYEN des GENERATIONS MASCULINES et FEMININES
en AGE de PREMIER MARIAGE dans six pays européens
Echelles des ordonnées logarithmiques de même module

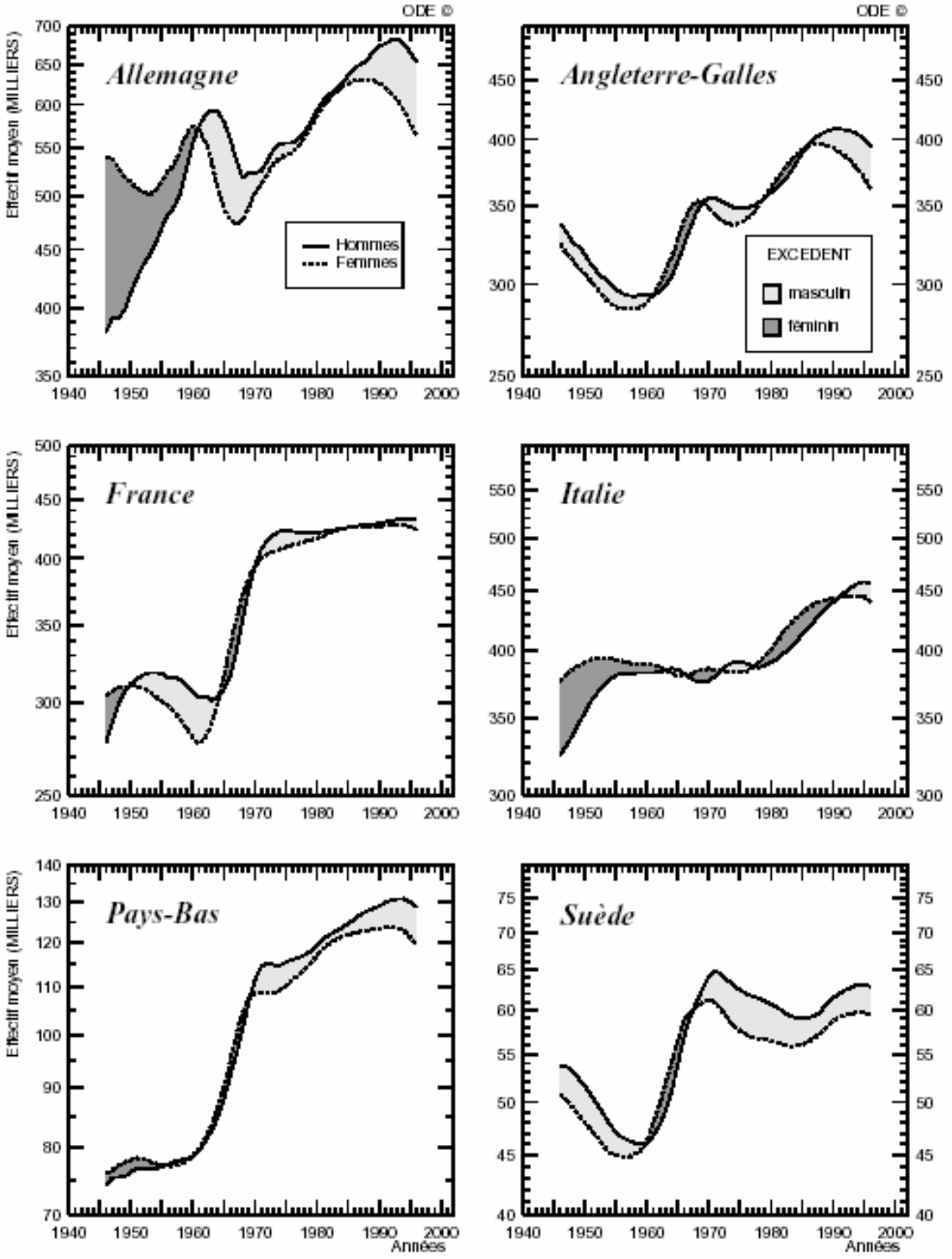


Figure 4 RAPPORT :
(1) en terme d'INDICATEUR CONJONCTUREL de PRIMO-NUPTIALITE de l'indicateur FEMININ à l'indicateur MASCULIN
(2) en terme d'EFFECTIF MOYEN des GENERATIONS en AGE de PREMIER MARIAGE de l'effectif MASCULIN à l'effectif FEMININ

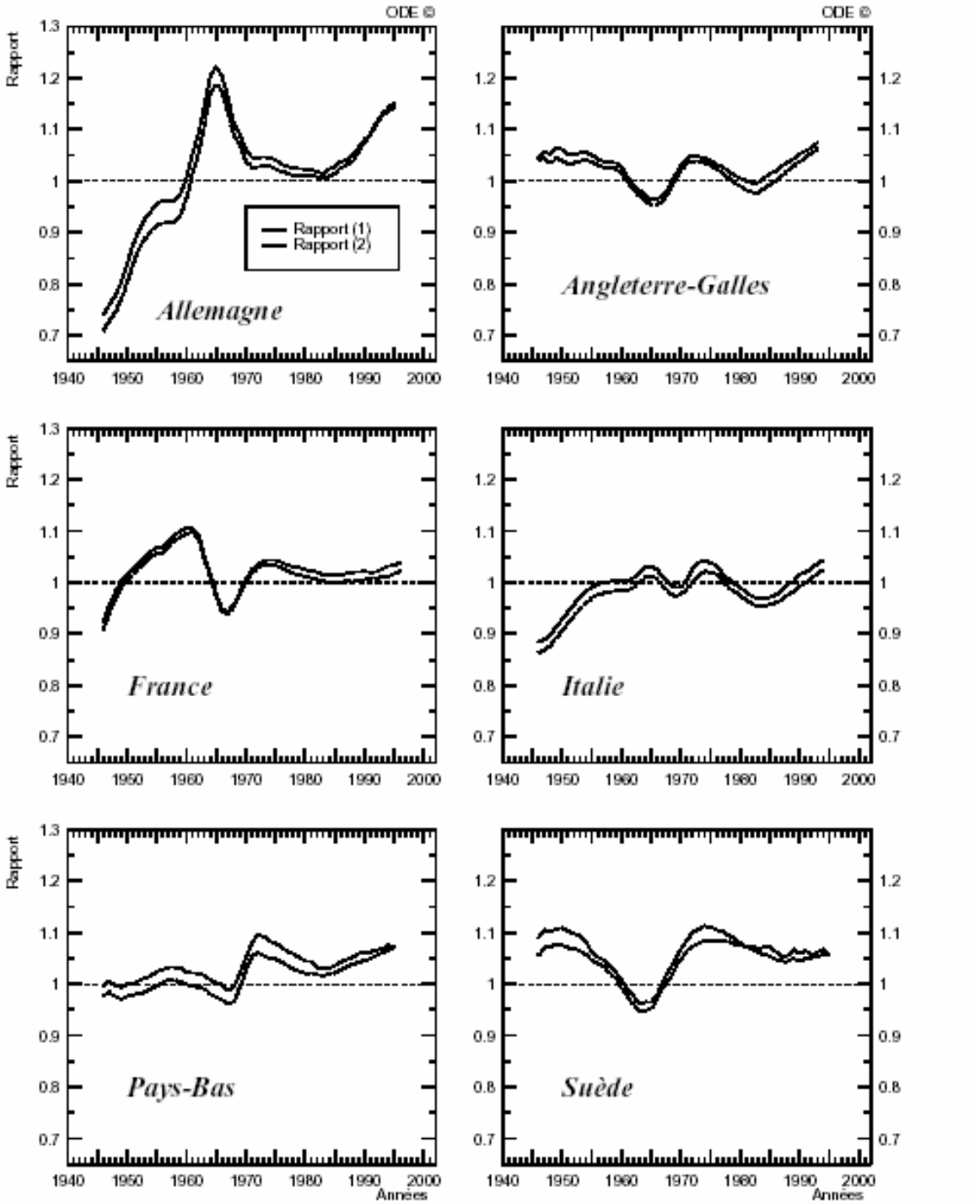
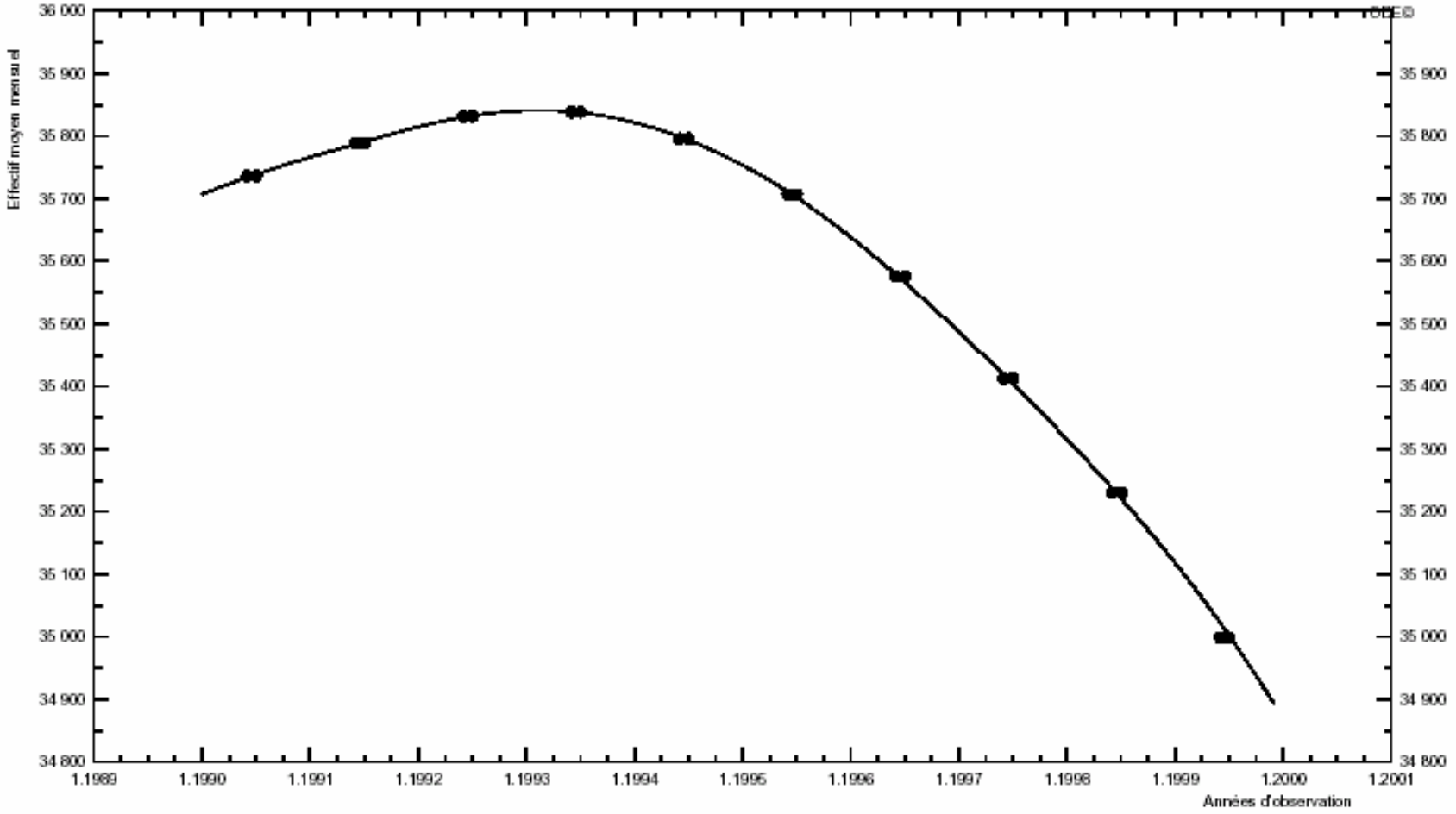


Figure 5 FRANCE, 1990-1999. EFFECTIF MOYEN des GENERATIONS FEMININES en AGE de FECONDITE
Valeurs annuelles divisées par 12 et rapportées à JUIN et JUILLET,
Valeurs MENSUELLES INTERPOLEES par polynômes de degré 5 sur 6 points consécutifs
dont on retient l'année centrale



En extrapolant (par un procédé quelconque) à l'année *n* ce rapport, qui est connu jusqu'à l'année $k = n - a$, puis en divisant l'estimation $\hat{G}^n$ par la valeur extrapolée à l'année *n* de ce rapport, on obtient une estimation améliorée de G^n.

L'extrapolation du rapport repose implicitement sur l'hypothèse de la stabilité des migrations et de la mortalité et, dans une moindre mesure, sur celle de la régularité de l'évolution des taux à âge égal.

Si on dispose d'une *projection* de population , au moins à l'horizon du 1er janvier *n*+1, on peut aussi utiliser les effectifs projetés P_{i-1}^n et P_i^{n+1}, c'est-à-dire estimer G^n par :

$$\hat{G}^n = \frac{\sum_{i=15}^{49} f_i^a \, \frac{P_{i-1}^n + P_i^{n+1}}{2}}{\sum_{i=15}^{49} f_i^a}$$

estimation qui peut à son tour être améliorée moyennant extrapolation, à l'année *n*, du rapport connu jusqu'à l'année $k = n-a$:

$$\frac{\hat{G}^k}{G^k} = \frac{\sum_{i=15}^{49} f_i^{k-(n-a)} \, \frac{P_{i-1}^k + P_i^{k+1}}{2}}{\sum_{i=15}^{49} f_i^{k-(n-a)}} \Big/ \frac{\sum_{i=15}^{49} f_i^m \, \frac{P_{i-1}^k + P_i^{k+1}}{2}}{\sum_{i=15}^{49} f_i^k}$$

6. L'estimation des indicateurs conjoncturels annuels et mensuels : comment convertir un nombre absolu d'événements en indicateur ?

Dès qu'on dispose d'une évaluation, même provisoire, d'un *nombre absolu* d'événements, on peut estimer l'indicateur conjoncturel qui lui correspond en divisant ce nombre absolu par la valeur de l'effectif moyen des générations soumises au risque. Ceci vaut aussi bien à l'échelle annuelle qu'à l'échelle mensuelle.

Toutefois, à l'échelle mensuelle, une opération préalable est nécessaire. Il convient en effet de corriger le nombre absolu observé de deux phénomènes perturbateurs : la *composition en jours* du mois (nombre de jours et, le cas échéant, nombres de lundis, de mardis, ..., de dimanches, si le phénomène étudié est soumis à une fluctuation hebdomadaire importante, comme il en va notamment en matière de mariages) et les *variations saisonnières* mensuelles.

On trouvera en annexe, deux tableaux donnant, pour la France, les résultats les plus récents dont on dispose à la date où nous écrivons (février 1998) sur la fécondité et la primo-nuptialité.

Dans les figures 6 et 7, on a représenté l'évolution mensuelle des indicateurs conjoncturels de fécondité et de primo-nuptialité en France, corrigés de la composition journalière du mois et des variations saisonnières. On a indiqué sur ces mêmes figures l'évolution des indicateurs conjoncturels *lissés* obtenus par application d'une formule de lissage, due à Jan Hoem (Université de Stockholm), qui fournit une valeur lissée jusqu'au dernier mois d'observation.

Pour estimer sur *longue période* l'évolution de l'effectif moyen des générations féminines d'âge fécond et par conséquent celle de l'indicateur conjoncturel de fécondité, il est *nécessaire* de disposer, *chaque année*, des effectifs de la population féminine par âge et il est *souhaitable* de disposer de taux de fécondité par âge qui ne soient pas trop obsolètes.

Cependant, même si les taux disponibles sont assez largement obsolètes, on peut quand même les utiliser, à moins qu'ils ne se rapportent à une année exceptionnelle ou qu'on veuille estimer l'indicateur conjoncturel d'une année elle-même exceptionnelle. La qualité de l'estimation obtenue pourra être appréciée en comparant, pour les années dont les taux par âge sont disponibles, l'indicateur conjoncturel estimé et l'indicateur conjoncturel observé.

On trouvera représentée dans la figure 8 l'évolution de l'indicateur conjoncturel de fécondité en Suisse, estimé à partir des nombres absolus de naissances totales et des effectifs féminins par âge au 1er janvier de chaque année depuis 1861, le jeu de taux par âge retenu étant invariablement celui observé en 1932. A partir de 1932, on dispose de la série annuelle des naissances par année d'âge de la mère, ce qui permet de calculer la valeur exacte de l'indicateur conjoncturel. On constate ainsi que, durant la période de 65 ans considérée (1932-1996), l'erreur maximale commise en utilisant le calendrier transversal de la fécondité de 1932 atteint 0,08 enfant pour une femme en 1968, l'erreur n'excédant 0,03 enfant pour une femme que de 1962 à 1975 et 0,05 enfant pour une femme que de 1964 à 1972. Rappelons que les décennies 1960 et 1970 correspondent, en Suisse comme dans le reste de l'Europe, à une époque où le calendrier de la fécondité était spécialement *précoce*, donc assez différent de celui de 1932 mais surtout où les générations en âge de fécondité étaient spécialement *inégales* du fait de l'arrivée progressive à l'âge de la maternité des générations du *baby-boom*.

7. L'indicateur conjoncturel mensuel de primo-nuptialité

Les nombres mensuels d'événements dont on dispose en matière de *mariages* se rapportent généralement à l'*ensemble* des mariages (quels que soient les âges des époux et quels que soient leurs états matrimoniaux antérieurement au mariage), tandis que la *primo-nuptialité*, par exemple masculine, ne concerne par convention que les *premiers* mariages et, au surplus, d'hommes qui avaient *moins de 50 ans révolus* au moment de leur mariage.

Aussi, lorsqu'on dispose d'un nombre total de mariages, est-il nécessaire d'estimer, selon le sexe, le nombre de *premiers mariages avant 50 ans* qui lui correspond. Ceci peut se faire moyennant extrapolation, et interpolation si on travaille à l'échelle mensuelle, de la série annuelle observée du rapport entre le nombre de premiers mariages avant 50 ans et le nombre de mariages totaux. Ces extrapolations et interpolations peuvent être réalisées de la même façon que les opérations analogues effectuées sur l'effectif moyen des générations soumises au risque.

8. L'indicateur conjoncturel mensuel de mortalité

L'indicateur conjoncturel mensuel de *fécondité*, pour le mois *m* de l'année *n*, s'obtient en divisant le nombre absolu mensuel NM^m de naissances observé le mois *m*, préalablement corrigé en NM^{m*} pour tenir compte de la composition journalière du mois et des variations saisonnières, par l'effectif moyen des générations féminines d'âge fécond établi pour le même mois *m*. C'est aussi l'indicateur conjoncturel qu'on *aurait* obtenu pour l'ensemble de l'année *n* si les effectifs de naissances selon l'âge de la mère avaient été égaux à $\frac{12\,NM^{m*}}{N^n} N_i^n$ au lieu de N_i^n.

Figure 6 FRANCE, 1977-1997
Evolution de l'INDICATEUR CONJONCTUREL mensuel de FECONDITE
Série corrigée des variations saisonnières
et série lissée par moyenne mobile Hoem sur 25 termes

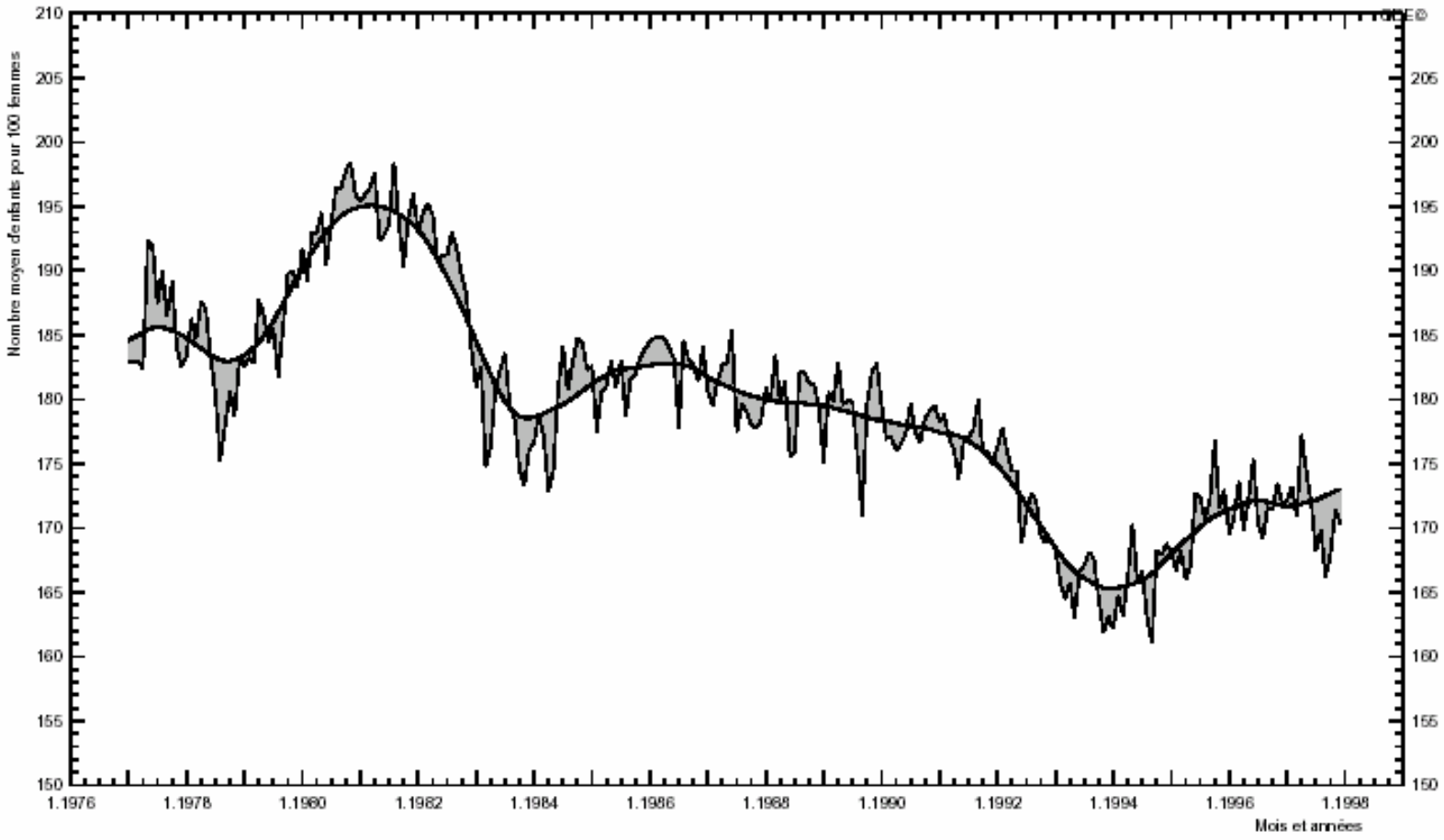
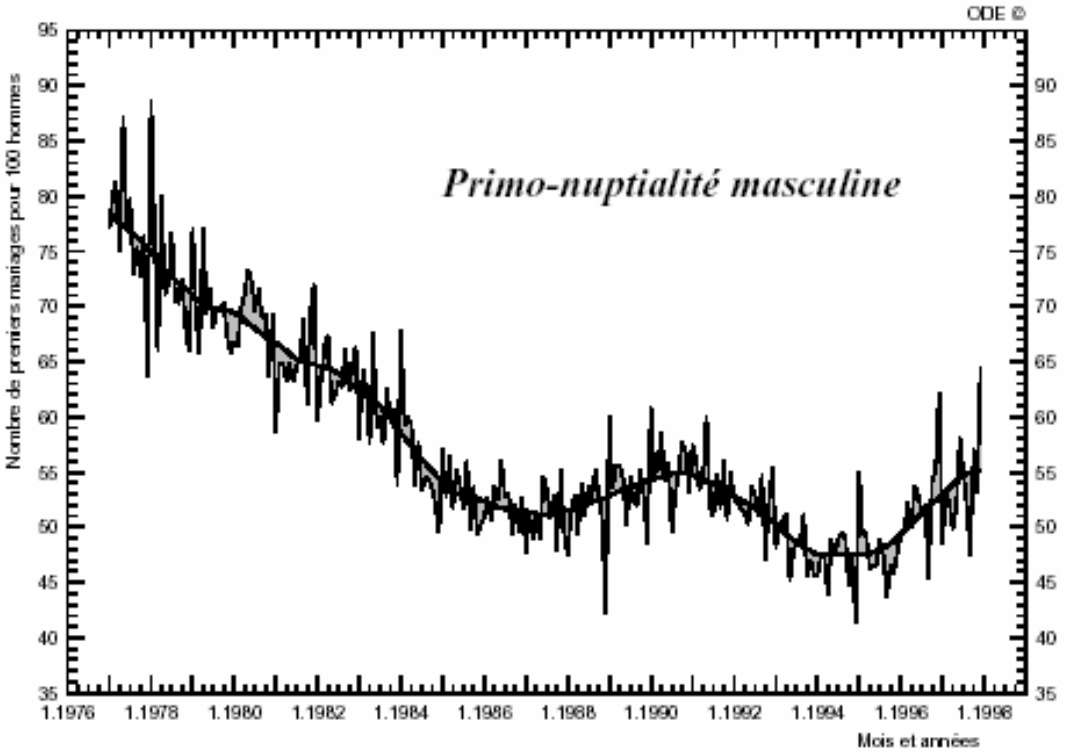
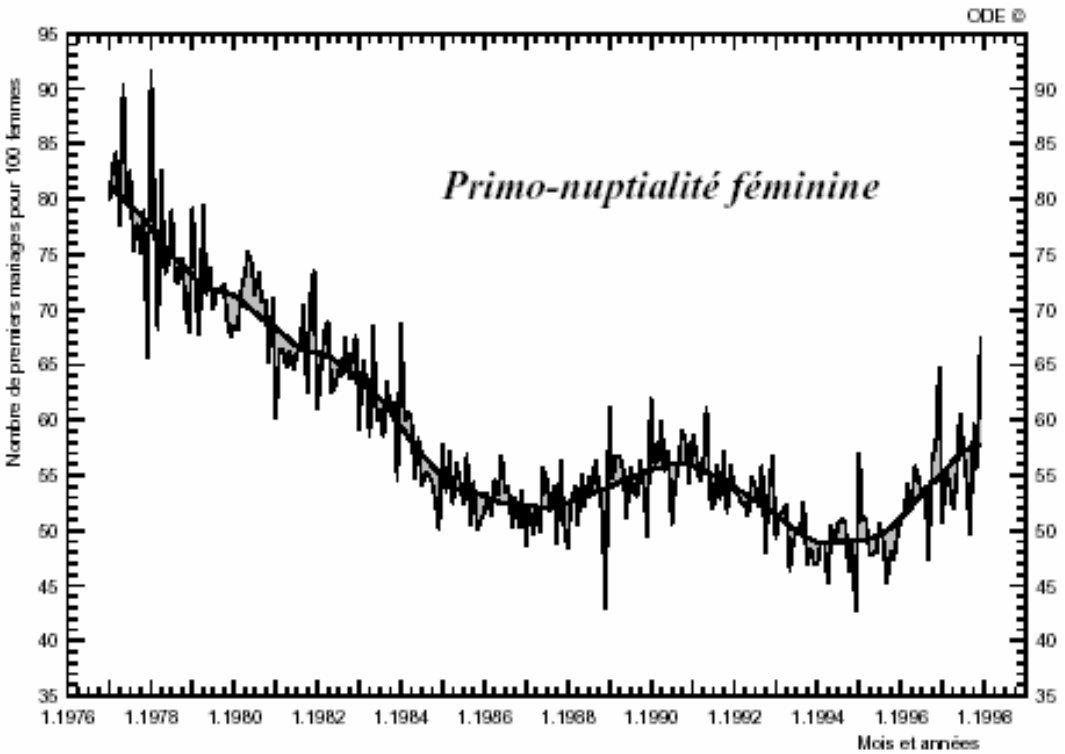


Figure 7 FRANCE, 1977-1997
Evolution de l'INDICATEUR CONJONCTUREL mensuel
de PRIMO-NUPTIALITE MASCULINE et FEMININE
Série corrigée des variations saisonnières,
puis lissée par moyenne mobile Hoem sur 25 termes





De façon analogue, on peut convenir de construire un indicateur conjoncturel mensuel de *mortalité*, exprimé en termes d'espérance de vie à la naissance masculine ou féminine, en déterminant l'espérance de vie qu'on *aurait* obtenue pour l'ensemble de l'année *n* si, au lieu des nombres de décès par sexe et âge réellement observés D_i^n, on avait enregistré ces mêmes nombres *multipliés* par $\frac{12 DM^{m*}}{D^n}$, expression où DM^{m*} désigne le nombre mensuel de décès du mois *m* corrigé de la composition journalière du mois et des variations saisonnières. On trouvera représentées dans la figure 9 les évolutions des indicateurs conjoncturels mensuels de mortalité masculine et féminine en France depuis vingt ans.

Du fait que le mouvement saisonnier des décès n'est pas indépendant du sexe et surtout de l'âge, l'indicateur conjoncturel mensuel de mortalité ainsi défini diffère de celui qu'on aurait établi si on avait disposé des nombres *mensuels* de décès par sexe et âge et construit une table de mortalité *mensuelle*. Il fournit cependant une description de l'évolution mensuelle qui est *cohérente* avec l'évolution de l'indicateur annuel (la moyenne des douze indicateurs mensuels est sensiblement l'indicateur annuel) et qui reproduit les variations conjoncturelles du nombre absolu mensuel. En particulier, les mois marqués par une épidémie de grippe, qui correspondent à un indicateur mensuel relativement *faible*, apparaissent avec netteté.

9. Conversion d'un nombre annuel de décès en espérance de vie à la naissance

On a vu plus haut la manière de convertir un nombre absolu annuel de *naissances* ou de *mariages* en l'indicateur conjoncturel correspondant (indicateurs conjoncturels de fécondité et de primo-nuptialité masculine et féminine) : on divise le nombre absolu d'événements, qu'on a préalablement exprimé en nombre de premiers mariages avant 50 ans dans le cas de la primo-nuptialité, par l'estimation de l'effectif moyen des générations soumises au risque.

La même question se pose de convertir un nombre annuel de *décès*, portant généralement sur l'*ensemble* des *deux* sexes, en les espérances de vie, masculine et féminine, à la naissance. Voici un procédé permettant d'opérer cette conversion.

Soit *a* l'année la plus récente pour laquelle on dispose de la table de mortalité par sexe et âge, table dont les espérances de vie à la naissance, masculine, féminine et deux sexes, sont désignées respectivement par Ev0H(*a*), Ev0F(*a*) et Ev0 (*a*).

Soit, de même, *b* (avec $b >= a+1$) l'année la plus récente pour laquelle on dispose des effectifs de population par sexe et âge au 1er janvier, *c* (souvent $c = a$) l'année la plus récente pour laquelle on dispose de la table de fécondité par âge et *m* (souvent $m = a$) l'année la plus récente pour laquelle on dispose des soldes migratoires par sexe et âge. Désignons par D^n l'évaluation du nombre absolu annuel de décès dont on dispose pour l'année *n* ($n > a$).

Figure 8 SUISSE, 1861-1996
Evolution de l'INDICATEUR CONJONCTUREL de FECONDITE
Données observées (1932-1996)
Données estimées (1861-1996) sur la base du calendrier de la fécondité observé en 1932

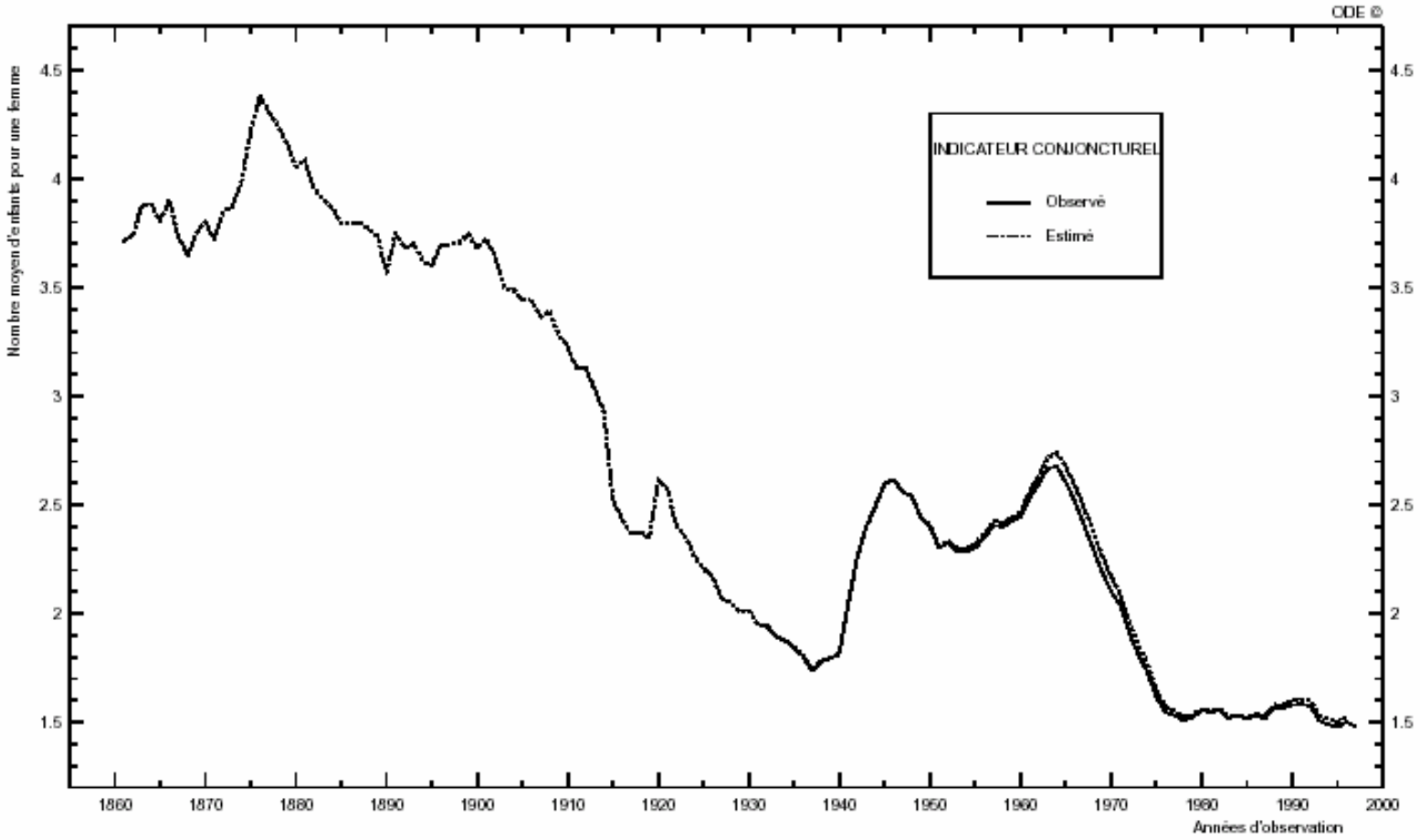
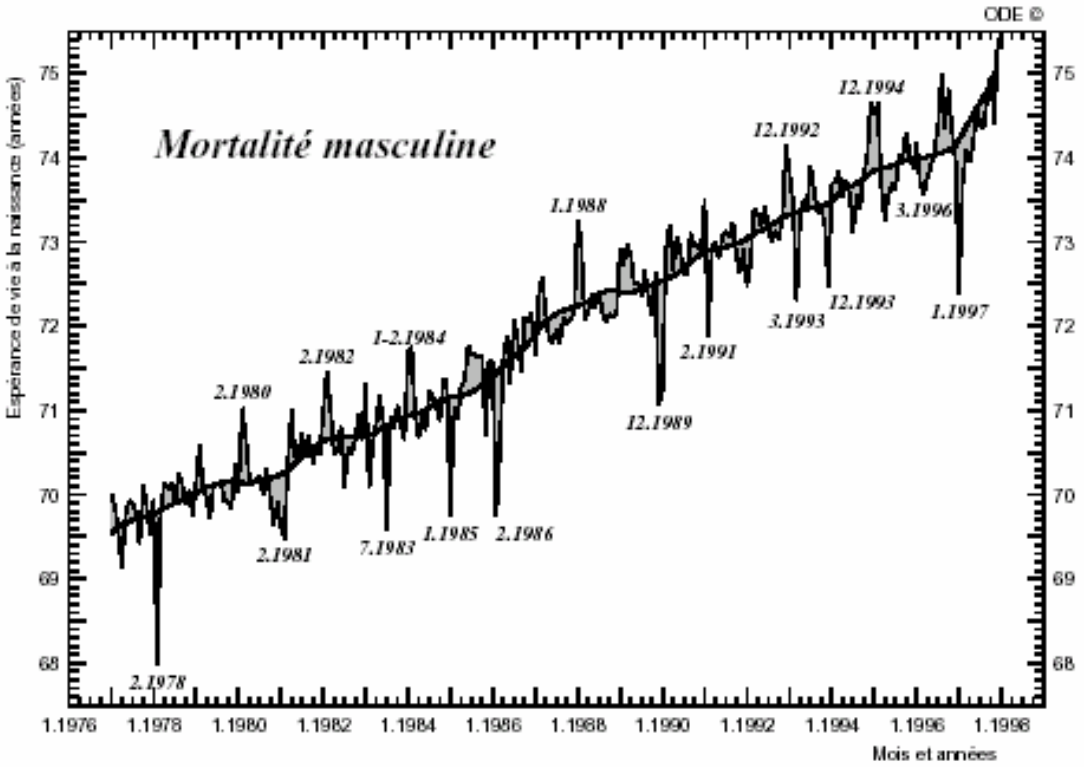
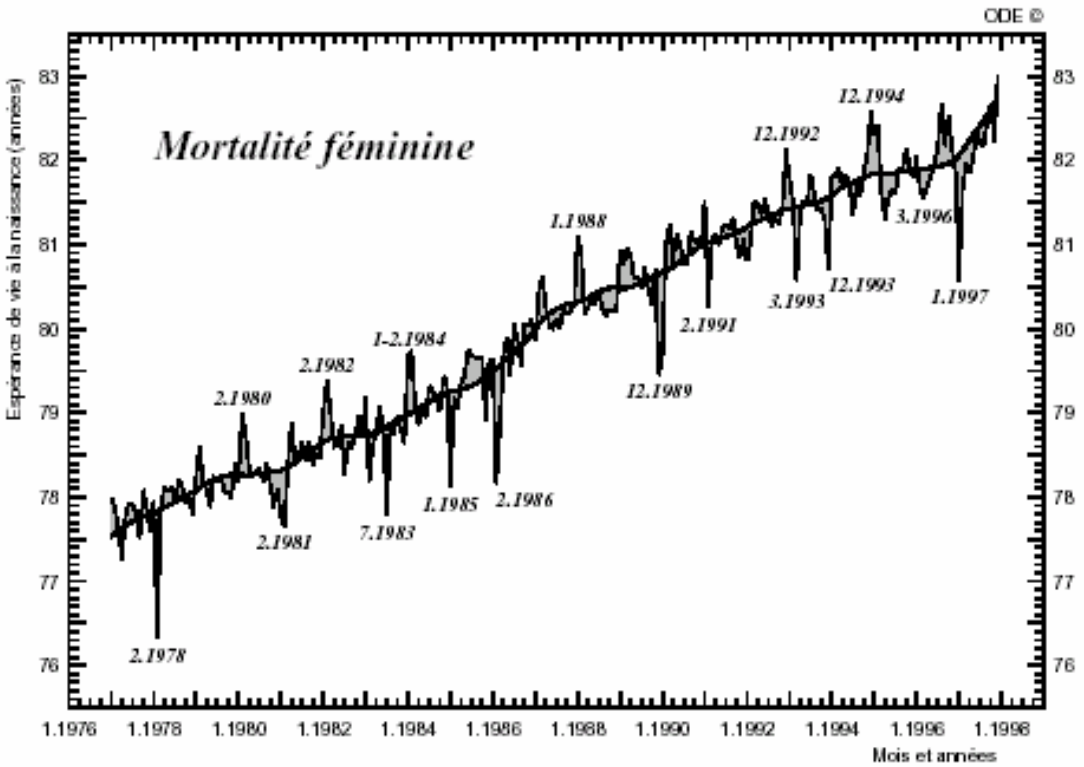


Figure 9 FRANCE, 1977-1997
Evolution de l'INDICATEUR CONJONCTUREL mensuel
de MORTALITE MASCULINE et FEMININE
Série corrigée des variations saisonnières
et série lissée par moyenne mobile Hoem sur 25 termes





On effectue une *série* de p+1 *projections* de population (p de l'ordre de 5 à 10) à mortalité *constante*, fécondité *constante* et soldes migratoires *constants* sous les hypothèses ci-après :

date de départ de la projection : 1er janvier k-(n-b), c'est-à-dire 1er janvier b-(n-k)
effectifs initiaux : ceux *réellement* observés par sexe et âge au 1er janvier b-(n-k)
date-horizon : 1er janvier k+1, soit n-b+1 bonds d'un an dans le temps
table de mortalité par sexe et âge de l'année a-(n-k),
table de fécondité par âge de l'année c-(n-k)
soldes migratoires par sexe et âge de l'année m-(n-k)

Ces projections de population sont réalisées successivement pour $k = b$-p+1, b-p+2, ..., b, puis pour $k = n$. De ces p+1 projections, on retient le nombre de décès *projeté* D^{k*} que l'on rapporte au nombre de décès *observé* D^k. Le gain d'espérance de vie à la naissance au cours de la période qui va de l'année a-n+k à l'année k, soit sur un intervalle de n-a années, est égal à Ev0H(k) - Ev0H(a-n+k) pour le sexe masculin, Ev0F(k) - Ev0F(a-n+k) pour le sexe féminin et Ev0 (k) - Ev0 (a-n+k) pour l'ensemble des deux sexes. La corrélation entre l'un ou l'autre de ces trois gains et le rapport du nombre de décès deux sexes projeté pour l'année k à mortalité constante (celle de l'année a-n+k) à celui observé, D^{k*}/D^k, est généralement étroite.

Sur le nuage, par exemple masculin, de p points d'abscisses D^{k*}/D^k et d'ordonnées Ev0H(k) - Ev0H(a-n+k), on détermine, par la méthode des moindres carrés, le paramètre a de la relation statistique (droite ajustée passant par le point de coorodonnées 0 et 1) :

$$D^{k*}/D^k = 1 + a\ [\text{Ev0H}(k) - \text{Ev0H}(a\text{-}n\text{+}k)]$$

unissant abscisses et ordonnées (Figure 10). Cette relation statistique est ensuite appliquée, pour $k = n$, au rapport D^{n*}/D^n, ce qui fournit Ev0H(n) - Ev0H(a) et donc l'estimation cherchée de Ev0H(n). On procède de la même façon pour le sexe féminin et pour l'ensemble des deux sexes.

On peut encore procéder de la même façon pour estimer les espérances de vie, masculine et féminine, non pas à la naissance, mais à un âge quelconque. La figure 11 est l'analogue de la figure 10 pour l'estimation de l'espérance de vie à 60 ans, sur la base du nombre absolu annuel de décès.

10. La signification d'un indicateur conjoncturel de fécondité ou de nuptialité

Le concept d'effectif moyen des générations soumises au risque, qu'il a été nécessaire d'adapter dans le cas de la mortalité, c'est-à-dire d'un événement *non renouvelable* dont les intensités sont mesurées par une série de *quotients* par âge, permet de préciser la signification d'un indicateur conjoncturel. Que l'événement soit renouvelable ou non, la démarche suivie pour apprécier la portée d'un nombre absolu d'événements est la même : elle consiste en une comparaison entre ce nombre absolu et un nombre de référence.

En matière de fécondité, on compare le nombre absolu des naissances, c'est-à-dire l'effectif de la génération née durant l'année, à l'effectif des générations adultes dont cette génération est issue. Cette comparaison est effectuée sur la base du sexe *féminin* : on rapporte l'effectif de la génération féminine née durant l'année à l'effectif moyen (pondéré) des générations féminines qui, cette année-là, ont l'âge d'avoir des enfants.

L'égalité entre ces deux effectifs, c'est-à-dire la valeur 1 de ce rapport, sert ainsi de *repère*, qu'on dénomme *remplacement* ou encore *strict remplacement*.

Etant donné que la proportion de filles à la naissance est invariablement de 100 *filles* pour 205 *naissances*, il est équivalent de considérer le nombre *total* de naissances de l'année et de prendre pour repère la valeur 2,05 du rapport – qui n'est alors autre que l'indicateur conjoncturel – c'est -à-dire de retenir 2,05 comme repère de l'indicateur conjoncturel de fécondité. Autrement dit, la valeur 2,05 enfants pour une femme, prise par l'indicateur conjoncturel de fécondité, signifie très exactement l'égalité entre le nombre de filles nées durant l'année et l'effectif moyen pondéré des diverses générations féminines qui, la même année, ont l'âge d'être mères.

On peut affiner très légèrement le repère en observant que la comparaison précédente porte, d'une part, sur des filles qui viennent de naître et, d'autre part, sur des femmes dont l'âge moyen est de l'ordre de 28 ans. En divisant la valeur-repère 2,05 par la proportion des filles qui atteindront à leur tour l'âge d'être mères, proportion de l'ordre de 0,985 si on se réfère aux tables de mortalité transversales actuelles, on aboutit à une nouvelle valeur-repère, égale à 2,08 et arrondie habituellement à 2,1 enfants pour une femme.

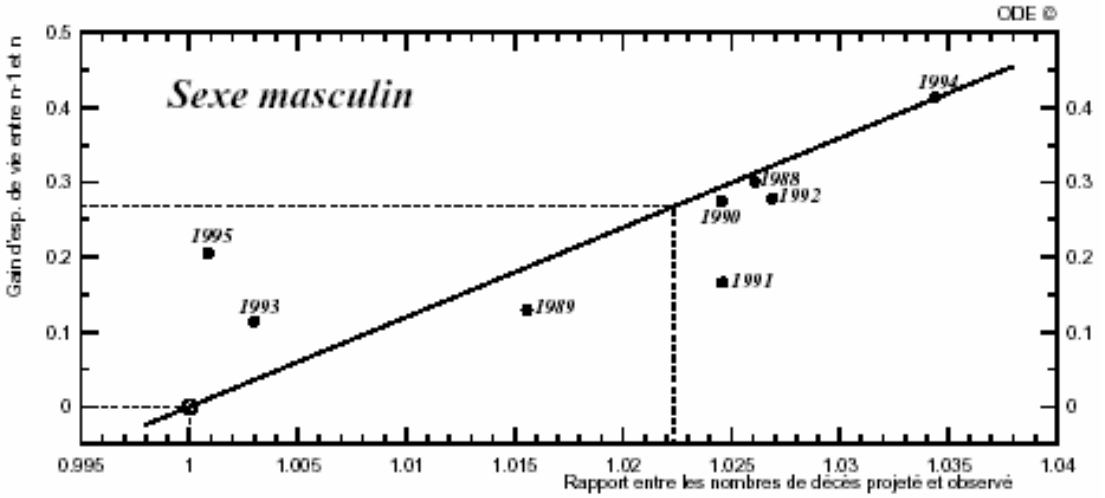
Dans ces conditions, la valeur 2,08 enfants pour une femme, prise par l'indicateur conjoncturel de fécondité, dont on dit qu'elle correspond au strict remplacement, signifie très exactement l'égalité entre l'effectif qui *sera*, en l'absence de migrations internationales, celui de la génération féminine née durant l'année, lorsqu'elle atteindra à son tour l'âge d'avoir des enfants, et l'effectif *moyen* des diverses générations féminines qui appartiennent *actuellement* au groupe d'âge fécond.

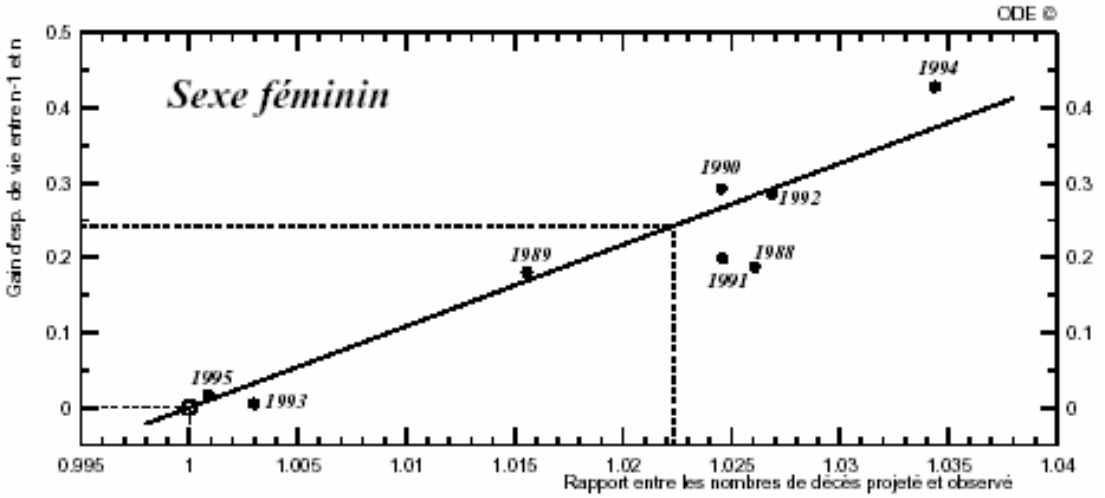
Plus généralement, le *rapport* de l'indicateur conjoncturel de fécondité à 2,08 est aussi le *rapport* entre l'effectif qui *sera*, en l'absence de migrations internationales, celui de la génération féminine née durant l'année, lorsqu'elle atteindra l'âge d'avoir des enfants, et l'effectif *moyen* des diverses générations féminines qui ont *actuellement* l'âge de la maternité.

Nous préférons cette définition de l'indicateur conjoncturel de fécondité à celle souvent donnée et que nous estimons critiquable, fondée sur l'artifice de la *cohorte fictive* : nombre moyen d'enfants auquel *parviendrait*, en fin de vie féconde, un ensemble de femmes qui, aux différents âges, *auraient* le même taux de fécondité que celui observé au même âge durant l'année – mais sur des générations réelles *différentes* –. En effet, cette définition repose implicitement sur l'hypothèse de la plausibilité de l'existence d'une telle génération. Or ce calcul peut fort bien être *irréaliste* dans la mesure où il est impossible d'imaginer qu'une génération *réelle* puisse avoir un tel comportement *tout au long de sa vie féconde*. Qu'on songe par exemple au cas de l'année 1916 en France : quel sens aurait le comportement d'une génération qui vivrait *toute* sa vie féconde dans les mêmes conditions, à âge égal, que celles qui prévalaient durant l'année 1916 où la quasi-totalité des jeunes hommes étaient au front ? En revanche, la valeur prise en 1916 par l'indicateur conjoncturel de fécondité, soit 1,21 enfant pour une femme, signifie que le nombre de filles nées en 1916 n'a atteint que 1,21/2,05 = 59% d'une classe d'âge féminine moyenne alors en âge d'avoir des enfants.

En matière de primo-nuptialité, par exemple masculine, l'indicateur conjoncturel est le rapport entre le nombre absolu de mariages d'hommes célibataires de moins de 50 ans célébrés durant l'année et l'effectif moyen (pondéré) des générations masculines, qui cette année-là ont l'âge du premier mariage. Quand l'indicateur conjoncturel de primo-nuptialité masculine vaut par exemple 0,6 premier mariage pour un homme, cela signifie que le nombre de premiers mariages célébrés avant 50 ans représente 60% d'une classe d'âge masculine moyenne en âge de premier mariage.

Figure 10 FRANCE, 1988-1996
Estimation de l'ESPERANCE de VIE à la naissance de l'année n selon le sexe connaissant les effectifs par sexe et âge jusqu'au 1er janvier n et les tables de fécondité, de mortalité et les soldes migratoires jusqu'à l'année n-1
En abscisses : rapport entre le nombre de décès projeté et le nombre observé
En ordonnées : gain d'espérance de vie à la naissance (en années) entre n-1 et n





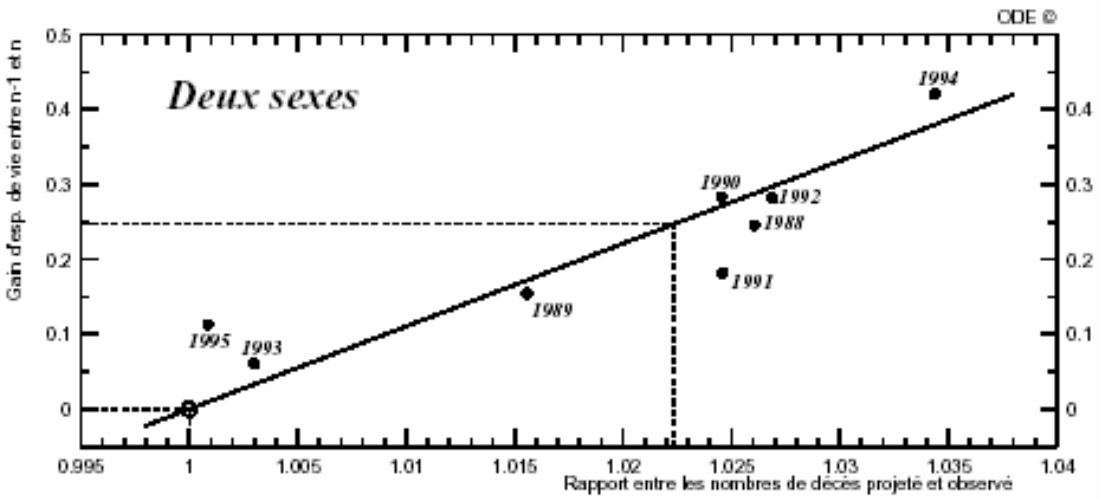
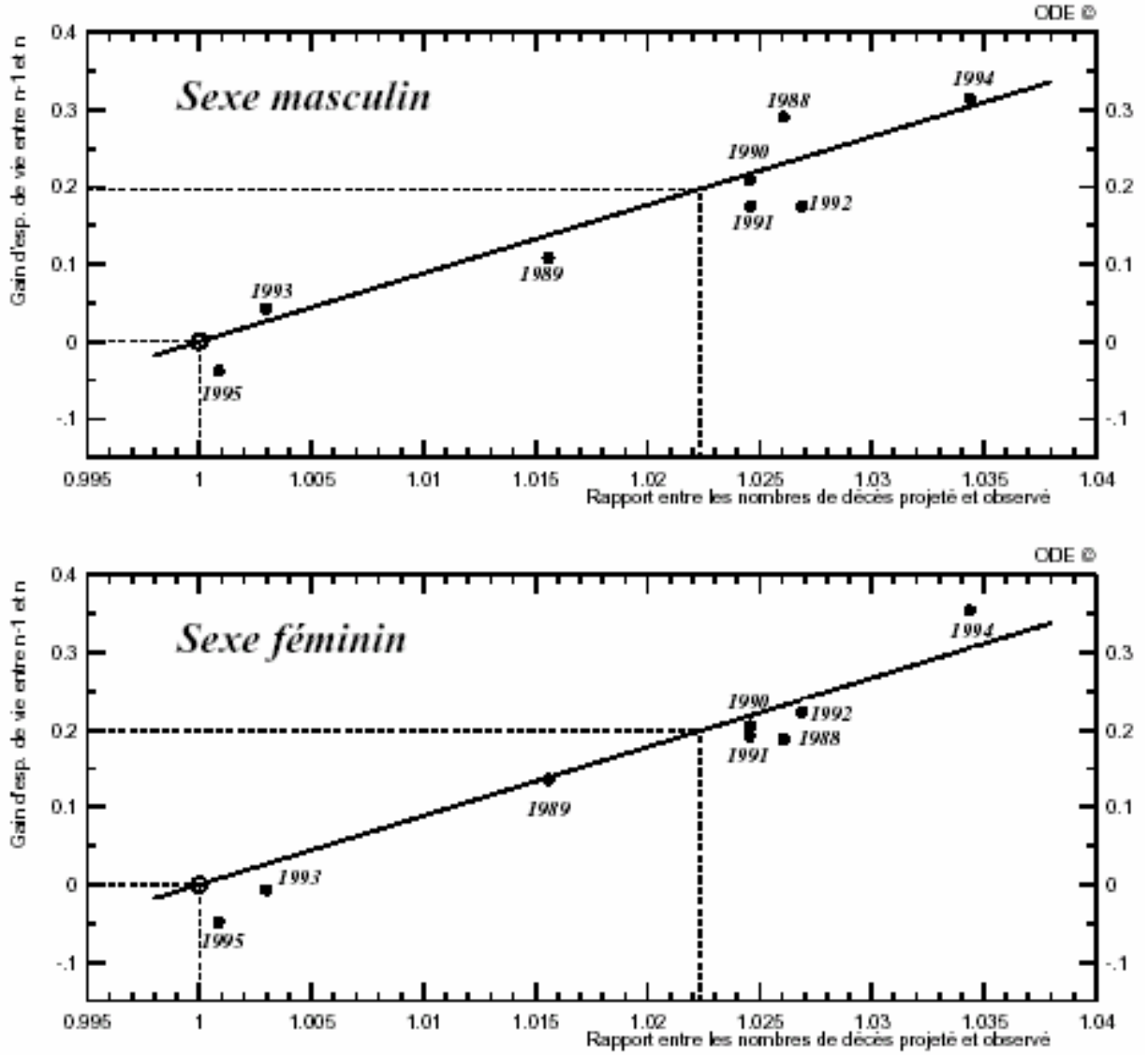


Figure 11 FRANCE, 1988-1996
Estimation de l'ESPERANCE de VIE à 60 ans de l'année n selon le sexe connaissant les effectifs par sexe et âge jusqu'au 1er janvier n et les tables de fécondité, de mortalité et les soldes migratoires jusqu'à l'année n-1
En abscisses : rapport entre le nombre de décès projeté et le nombre observé
En ordonnées : gain d'espérance de vie à 60 ans (en années) entre n-1 et n



11. Flux annuel d'événements et intensité/calendrier du phénomène

Revenons au nombre annuel de naissances. Nous avons vu plus haut que ce qui en conditionne la valeur l'année *n*, c'est *quasi*-exclusivement l'indicateur conjoncturel I^n de cette même année *n*, puisque l'effectif moyen des générations féminines d'âge fécond de l'année *n* est *quasi*-entièrement déterminé par l'évolution démographique au cours des années *antérieures* à *n*. Le *moteur* du nombre absolu des naissances n'est donc pas directement le niveau de fécondité des générations qui, l'année *n*, ont l'âge d'être mères, c'est-à-dire la *descendance finale moyenne* des générations qui, l'année *n*, appartiennent aux divers âges féconds. Observons que cette descendance finale moyenne peut être prise comme égale à la descendance finale $DF\left(n-\bar{x}^n\right)$ de la génération née en *n*-$\bar{x}^n$, où $\bar{x}^n$ est l'âge moyen (transversal) à la maternité observé l'année *n*, compte tenu du fait que la descendance finale varie en général *lentement* d'une génération à l'autre.

Si on veut relier le nombre absolu des naissances observé l'année *n* au niveau de fécondité de la génération née en *n* - $\bar{x}^n$, il faut faire intervenir, comme facteur en quelque sorte *perturbateur*, le rapport entre I^n et $DF\left(n-\bar{x}^n\right)$. Or ce rapport varie lui-même de façon complexe, sous l'effet des variations du calendrier transversal de la fécondité au fil des années successives.

On peut montrer[67] que sous les hypothèses très particulières suivantes :

- l'indicateur conjoncturel I^n des années successives est *invariant*
- la distribution de l'âge transversal à la maternité varie de telle sorte que tous ses moments *centrés* demeurent cependant *invariants*
- l'âge moyen transversal à la maternité $\bar{x}^n$ varie *linéairement* avec *n*

la descendance finale est *invariante* et, en désignant par $\bar{x}'$ la dérivée *constante* de $\bar{x}^n$ par rapport à *n*, le rapport de l'indicateur à la descendance finale est égal à :

$$\frac{I}{DF} = 1 - \bar{x}'$$

En d'autres termes, l'indicateur conjoncturel est la descendance finale d'un régime de fécondité *rigoureusement invariable* (taux de fécondité invariants à âge égal) qui conduirait, avec les *mêmes* effectifs féminins par âge F_i^n que ceux présents l'année *n* et avec les *mêmes* taux f_i^n que ceux enregistrés l'année *n*, au *même* nombre de naissances N^n que celui observé. Aux époques où le calendrier de la fécondité évolue rapidement, l'indicateur conjoncturel peut s'écarter notablement de la descendance finale moyenne des générations alors en âge de fécondité et une approximation, d'ailleurs assez grossière, de leur rapport est suggérée par la relation ci-dessus (indicateur *excédant* la descendance finale moyenne de l'ordre de 10% lorsque l'âge moyen à la maternité *s'abaisse* au rythme de 0,1 an par an, inversement descendance finale moyenne excédant l'indicateur de l'ordre de 10% lorsque l'âge moyen à la maternité *s'élève* au rythme de 0,1 an par an).

Ne perdons cependant pas de vue que ce qui importe en matière de fonctionnement de la machinerie démographique, ce n'est pas le *niveau de fécondité* (mesuré par la descendance finale) des générations qui se trouvent à l'époque considérée appartenir aux âges féconds, mais le *nombre absolu des naissances*. Or le nombre absolu des naissances dépend quasi-exclusivement, les effectifs féminins en âge de fécondité étant *donnés*, de l'indicateur conjoncturel. Même s'il est vrai que celui-ci constitue une image imparfaite de la

[67] Voir par exemple G. CALOT, Relationships between cohort and period demographic indicators, *Population*, An English selection, 5, 1993, 183-222.

descendance finale des générations qui sont alors d'âge fécond, en raison de l'effet perturbateur des variations du calendrier, c'est lui et non la descendance finale, qui détermine le nombre absolu des naissances.

Observons par ailleurs que ce que nous avons appelé indicateur conjoncturel est une mesure d'*intensité* dans le cas de la fécondité et de la primo-nuptialité, dont les événements sont (ou sont traités comme) *renouvelables*, tandis qu'en matière de mortalité, dont les événements sont *non renouvelables*, l'indicateur conjoncturel s'exprime en termes d'espérance de vie à la naissance, caractéristique non pas d'intensité mais de *calendrier*. C'est qu'en matière de mortalité, l'intensité ne soulève aucune question ; elle est *invariablement* égale à l'unité : à toute époque et dans toutes les générations, les humains sont *tous* mortels. Dans le cas des événements renouvelables, le flux annuel d'événements est en premier lieu sensible aux variations de l'intensité transversale ; dans le cas des événements non renouvelables dont l'intensité est égale à l'unité (événements qualifiés de *fatals* dans la littérature), il est en premier lieu sensible aux variations du calendrier transversal. Il n'en irait pas de même si on traitait la primo-nuptialité comme produisant des événements non renouvelables : le flux annuel serait alors sensible *à la fois* aux variations de l'intensité transversale et à celles du calendrier transversal : la corrélation sur laquelle nous nous sommes appuyés pour estimer l'espérance de vie à partir du nombre annuel de décès n'aurait pas de sens en matière de primo-nuptialité.

Index of indicators for which the methodology is provided

Population

Mean population: p. 36

Estimate of the population on 1 January based on a cohort on any date: p. 23

Mean cohort of generations subject to the risk: p. 66 and 70

Fertility

Fertility rate by age:

- age reached: p. 50
- age completed: p. 42
- five-year age: p. 18

Total fertility rate: p. 66

Completed fertility: p. 66

Mean transversal age at childbearing:

- for all orders: p. 34 and 67
- order 1: p. 61
- order 2: p. 61
- order 3: p. 61
- order 4 and over: p. 61

Mean longitudinal age at childbearing:

- for all orders: p. 34 and 67
- order 1: p. 61
- order 2: p. 61
- order 3: p. 61
- order 4 and over: p. 61

Definitive infertility in the generations: p. 68

Proportion of women, in the generations, with:

- 1 child: p. 68
- 2 children: p. 68
- 3 children: p. 68
- 4 children or more: p. 68

Parity progression ratio: p. 68

Nuptiality

First marriage rate by age and sex:

- age reached: p. 50
- age completed: p. 42
- five-year age: p. 18

Total first marriage rate: p. 66

Intensity of the first marriage in the generations: p. 66

Mean transversal age at 1st marriage: p. 34 and 67

Mean longitudinal age at 1st marriage: p. 34 and 67

Divorce rate

Total divorce rate: p. 62

Proportion of marriages dissolved by divorce in the cohorts: p. 62

Mean age of marriage at the time of divorce: p. 62

Mean age of marriage at the time of divorce in the cohorts: p. 62

Mortality

Probability of dying by age and sex: p. 53

Mortality rate: p. 53

Life expectancies by age and sex: p. 67