SURVIVAL MODELS IN ACTUARIAL MATHEMATICS: FROM HALLEY TO LONGEVITY RISK *

Ermanno Pitacco

Dipartimento di Matematica Applicata "B. de Finetti", University of Trieste

p.le Europa, 1 - I 34127 Trieste (Italy)

phone: +39 040 5587070

fax: +39 040 54209

email: ermanno.pitacco@econ.units.it

Abstract

Several contributions, which can be considered as landmarks in the evolution of survival modelling in actuarial mathematics, are focussed. The following topics are concerned: the development from age-discrete to age-continuous modelling, from single-decrement to multiple-decrement models and multistate models, from population homogeneity to population heterogeneity, from static mortality to the analysis of mortality trends and the consequent need for mortality projections. Finally, the paper places a special emphasis on some issues concerning actuarial modelling in a dynamic context. So, the topic of longevity risk is dealt with, aiming in particular to stress its peculiarity in the context of the mortality risks borne by an insurer (or a pension plan).

Keywords

Early actuarial models, Mortality laws, Multistate models, Heterogeneity, Mortality trends, Projections, Longevity risk

* Invited lecture, 7th International Congress INSURANCE: MATHEMATICS & ECONOMICS, ISFA, Lyon, June 2003

1. INTRODUCTION

The aims of this paper are to describe some aspects of the development of survival modelling in actuarial mathematics and to focus on a "new" type of risk affecting the management of life insurance and annuities portfolios, i.e. the longevity risk. In no way this paper aims at providing the reader with a "history" of survival modelling. First, we will simply focus on some scientific (and practical) contributions which, in our opinion, can be considered as landmarks in the evolution of survival modelling. In particular, we have tried to find who first thought of the several concepts and tools used in the actuarial field, pertaining to the area of survival modelling. Secondly, we will concentrate our attention on some recent works that show present trends in actuarial research complying with problems recently arisen in the context of life insurance and pension practice.

Survival models have played a central role throughout the whole history of life insurance and pension plans management. Although in many insurance products the investment component is of great importance, in the last decades concern for mortality issues has grown because of mortality trends experienced in many countries. Recent models for mortality projections and research work dealing with the uncertainty in future trends and the relevant actuarial evidence clearly show this interest. For this reason, the present paper places a special emphasis on some issues concerning survival modelling in a dynamic context.

The paper is organized as follows. In Section 2 the early actuarial models, proposed in the second half of the 17th century, are presented and the relevant features are discussed. Section 3 deals with the development of survival modelling and its applications to life insurance mathematics, and in particular the following aspects are analyzed: age-discrete vs age-continuous modelling, single decrements vs multiple decrements models, population homogeneity vs population heterogeneity. Mortality trends and the consequent need for mortality projections are discussed in Section 4, in which the concept of dynamic mortality modelling is sketched. Section 5 deals with the topic of longevity risk, aiming in particular to stress its peculiarity in the context of the mortality risks borne by an insurer (or a pension plan). Finally, some remarks in Section 6 conclude the paper.

The paper also aims at providing some guidelines for further reading, especially in the field of mortality forecasts. Several bibliographic references are given throughout the whole paper. Now we only mention some works which the reader interested in the history of actuarial mathematics and the applications of demography to life insurance should refer to. The paper by Haberman (1996) presents a very interesting introduction to the history of actuarial science up to 1919. From the discussion of over 100 contributions, representing landmarks in the actuarial context, the most important features of the development of theory and practice in this field clearly emerge. To a large extent, Section 3 of the present paper is based on this work. The paper by Haberman (1996) is a revised version of an essay introducing a collection of landmark works (Haberman and Sibbett, 1995).

Daw (1979) addresses the origins of multiple decrement modelling. Starting from the pioneering work of Daniel Bernoulli, in which the effects of inoculation against smallpox are analyzed, the paper illustrates the early contributions concerning the construction of multiple decrement tables and related single decrement tables. The paper by Seal (1977) presents the history of multiple decrement models, from Bernoulli's work up to the contributions of the 1960's.

Hald (1987) specifically deals with the early actuarial models, proposed in the second half of the 17th century for the evaluation of annuities on lives. The paper by Kopf (1926) deals with the early history of life annuities, focussing in particular on seminal contributions to the relevant actuarial structure.

The early attempts to consider mortality trends and the relevant actuarial evidence are described by Cramér and Wold (1935). This paper also presents and discusses early projection models. The book by Smith and Keyfitz (1977) collects important contributions to the mathematical development of demography; several contributions, especially in the fields of life tables and parametric curve fitting, are of great interest to actuaries, being related to the development of life insurance mathematics.

2. THE EARLY ACTUARIAL MODELS

2.1 The actuarial value of a life annuity

Let us consider an immediate life annuity of 1 per annum payable in arrear; let x denote the present age of the annuitant and a_x the expected present value (i.e. the actuarial value) of the annuity. The actuarial value can be calculated via the following formulae:

$$a_x = \sum_{t=1}^{+\infty} a_{\overline{t|}} \; {}_{t/1}q_x \tag{2.1}$$

$$a_x = \sum_{t=1}^{+\infty} (1+i)^{-t} \; {}_t p_x \tag{2.2}$$

where

$$a_{\overline{t|}} = \sum_{h=1}^{t} (1+i)^{-h} \tag{2.3}$$

is the present value of an annuity certain, i is the rate of interest, ${}_{t/1}q_x = \frac{l_{x+t} - l_{x+t+1}}{l_x}$ and ${}_t p_x = \frac{l_{x+t}}{l_x}$, with l_y denoting the (expected) number of survivors at age y in a given life table, assumed as a survival model. Using (2.3) in formula (2.1) and changing the order of summation one finds formula (2.2).

These formulae reflect, in a modern form, calculation procedures proposed in the second half of the 17th century. Actually, formula (2.1) generalizes the calculation procedure proposed in 1671 by the Dutch prime minister Jan de Witt, while formula (2.2) was proposed in 1693 by Edmund Halley, the famous astronomer. It is worth noting that formula (2.2) is computationally more straightforward, whereas formula (2.1) is much more interesting for further developments. Indeed formula (2.1) represents the expectation of the random present value $a_{\lceil K_x \rceil}$, where K_x denotes, according to the modern notation, the random curtate residual lifetime of a person aged x. Referring to this random variable, de Witt's method can be easily adapted to calculate higher moments, e.g. the variance of $a_{\lceil K_x \rceil}$, as pointed out by Haberman (1996).

2.2 Some features of the early model

The following features of the survival model used for evaluating life annuities should be pointed out. In a modern perspective, the model was: (a) deterministic; (b) age-discrete; (c) single decrement: (d) (implicitly) assuming homogeneity; (e) (implicitly) static. Early actuarial models for insurance products other than life annuities had similar features. This was the case, for instance, of the model proposed by James Dodson in 1755, for calculating level premiums in whole life assurance (see Haberman, 1996). Some comments about these aspects follow.

(a) Although de Witt's formula refers to the expected value of a random variable, the only language available in the latter half of the 17th century for describing probability models was the language from games of chance, as pointed out by Hald (1987). Actuarial models for life insurance have been explicitly proposed in terms of random variables just in the 1950's; de Finetti (1950, 1957) and Sverdrup (1952) first defined the random present value, Y, of insurance benefits as a function of the random residual lifetime T_x. So, for a whole life assurance of unitary amount we have $Y = (1+i)^{-T_x}$, for an endowment $Y = (1+i)^{-\min(T_x,n)}$ (with n denoting the term), for a continuous annuity of 1 per annum $Y = \bar{a}_{\lceil T_x \rceil}$ (with $\bar{a}_{\lceil t \rceil}$ denoting the present value of a continuous annuity paid up to time t, $a_{\lceil t \rceil} = \int_0^t e^{-\delta u} du$ and δ the instantaneous force of interest), etc. Apart from any language aspect, it is important to remark that the early survival model, albeit referring to random variables, did not allow for the riskiness inherent in insurance contracts, and hence can be considered as "deterministic".

(b) Halley's formula for the evaluation of life annuities constitutes one of the implementations of his life table, constructed from observed numbers of deaths in Breslau, whereas de Witt's life table was hypothetical. In both cases, since the proposed formulae explicitly refer to survival tables, it is quite natural that the adopted model is an age-discrete one. An important step towards age-continuous modelling follows from the early mortality "laws" originated from the fitting of mathematical formulae to mortality data.

(c) The type of benefits concerned in the early actuarial models, i.e. life annuity benefits (and assurances as well), naturally lead to a single-decrement setting. In the actuarial field, resorting to multiple decrement models follows the need to evaluate benefits depending on health status.

(d) Heterogeneity in respect of mortality is one of the most important issues in both survival modelling and actuarial practice. Although the early actuarial model did not allow for heterogeneity in populations, the problem of adverse selection was carefully considered at that time. As pointed out by Hald (1987), de Witt stressed that the nominee of an annuity contract usually is a person in "full health, and with a manifest likelihood of prolonged existence", whence a low mortality in the initial annuity period (at least) follows.

(e) It was not until the construction of a long series of mortality observations that trends in mortality clearly emerged and hence the concept of dynamic mortality was achieved, namely at the beginning of the 20th century. At present, allowing for mortality trends is one of the most important issues in actuarial modelling, especially when life annuities and other living benefits are concerned.

The contributions underpinning the early survival models were progressed further, and actuarial models as well. Development of survival modelling required a lot of work, involving actuarial science, probability theory, demography, medical statistics, etc. In recent times, numerical approaches to actuarial problems gained effectiveness thanks to the availability of high speed computers, so paving the way to a new "computational" actuarial mathematics, also based on stochastic simulation procedures. It is worth noting that, unfortunately, many interesting results were ignored for decades and practically forgotten, before being rediscovered and finally implemented. Moreover, a number of results in demography were ignored by actuaries and vice versa.

In Section 3 we present some aspects of the progression of survival modelling and the related development of actuarial models and implementations, referring to points (a) to (d); in Section 4 the development of a dynamic approach to mortality (point (e)) is specifically dealt with.

Further aspects should be analyzed to achieve a more complete description of the development of survival models and relevant applications to life insurance and pension problems. In particular, the availability of reliable life tables constitutes one of the bases for life insurance actuarial calculations, and hence improvements in table construction techniques represent important contributions also to progress in the actuarial field, especially when the analysis of mortality among insureds, annuitants and pensioners is concerned. For the same reason, research work concerning graduation techniques also plays a central role in the framework of actuarial science. Finally, studies concerning the comparison of actual and expected deaths should not be neglected. However, we disregard these fields of research because contributions therein obtained usually do not affect the "structure" of actuarial models. The reader interested in contributions belonging to these areas should refer to Benjamin and Pollard (1993) and Haberman (1996), also for bibliographic references.

3. THE DEVELOPMENT OF SURVIVAL MODELLING (AND RELATED ISSUES)

3.1 From deterministic to stochastic

Progression towards a stochastic approach to life insurance mathematics started at the end of the 18th century. In 1786 Johannes Tetens first addressed the analysis of mortality risk inherent in an insurance portfolio. The evidence of the role of $\sqrt{n}$ in determining the riskiness of a portfolio, where n denotes the number of policies in the portfolio itself, can be traced to Tetens' contribution. In particular, as pointed out by Haberman (1996), Tetens showed that the risk in absolute terms increases as the portfolio size n increases, whereas the risk in respect of each insured decreases in proportion to $\sqrt{n}$. In a modern perspective, Tetens' ideas constitute a pioneering contribution to individual risk theory.

Notable contributions to the development of a stochastic framework follow Tetens' ideas, throughout the 19th century. In 1859 Carl Bremiker attacked the problem of assessing the risk arising from mortality random fluctuations in a life assurance portfolio. In particular, he addressed the total claim distribution and proposed risk measures based on the assumption that the insured's lifetime is a random variable. For example, referring to a single premium whole life assurance Bremiker considered as a risk measure the variance of the random profit

$$\sigma^2 = \sum_{t=0}^{+\infty} {}_{t/1}q_x \left(\bar{A}_x - v^{t+\frac{1}{2}} \right)^2 \tag{3.1}$$

where $\bar{A}_x$ denotes the single premium and q a realistic assessment of the mortality rate.

As Haberman (1996) notes, Bremiker did not allow for other sources of random deviations in the results. In particular, the idea of a random financial result will be achieved after the seminal contribution of Louis Bachelier in 1900, concerning the stochastic modelling of investment problems. It is worth noting that stochastic finance will enter much later the life insurance actuarial context, in particular thanks to the work of F. M. Redington, dated 1952, addressing the principles of life office valuation.

Random mortality fluctuations constitute a "pooling" risk, i.e. a risk whose severity, conveniently assessed, decreases as the portfolio size n increases, as shown by Tetens. It should be stressed that the early contributions to a stochastic approach to life insurance mathematics addressed this type of risk only. Risks of systematic deviations, and thus “non-pooling” risks (parameter risk and model risk, in a modern perspective), will enter life insurance mathematics many years later, as we will see in Section 5.

A remarkable contribution to stochastic mortality in actuarial models came in 1869 from the work of Karl Hattendorff, who addressed the annual random losses of a life insurance policy. The annual random loss in year t, Λ_t, is defined as follows

$$\Lambda_t = \begin{cases} C\ (1+i)^{-1} - (V_{t-1} + P) & \text{if } \ t-1 \le T_x < t \\ V_t\ (1+i)^{-1} - (V_{t-1} + P) & \text{if } \ T_x \ge t \end{cases} \tag{3.2}$$

for $T_x \ge t-1$, whilst $\Lambda_t = 0$ otherwise; C denotes the sum assured, P the annual level premium, V_t the mathematical reserve at time t. Hattendorff's theorem states that:

$$E(\Lambda_t) = 0; \quad t = 1, 2, \dots \tag{3.3a}$$

$$cov(\Lambda_t, \Lambda_s) = 0; \quad t \neq s \tag{3.3b}$$

with the proviso that (3.3) are evaluated with the technical basis used for P and V_t. For the total loss $L = \sum_t \Lambda_t (1+i)^{-(t-1)}$ we have then:

$$var(L) = \sum_t var(\Lambda_t)(1+i)^{-2(t-1)} \tag{3.3c}$$

These results provide the actuary with a practicable formula for assessing the risk inherent in a policy, in terms of the variance of the total loss.

Both Bremiker and Hattendorff also focussed on the problem of facing adverse fluctuations. The need for an appropriate fund and, respectively, for a convenient safety loading of premiums emerged in their contributions. So these works can be considered as early landmarks in what we now call risk theory.

3.2 From age-discrete to age-continuous

Moving actuarial modelling from an age-discrete to an age-continuous context requires a more mature probabilistic framework, not strictly related with games of chance. In particular, as mortality represented the only random component in early actuarial models (and for a long time thereafter, as well), a very important step towards age-continuous modelling was the fitting of mathematical formulae to the observed data, in other words the step from mortality tables to mortality “laws”.

3.2.1 From table-based to law-based modelling

Probably, the first mathematical formula expressing the number of survivors as a function of the attained age x was proposed in 1725 by Abraham De Moivre, who suggested

$$l_x \;=\; k\left(1-\frac{x}{86}\right) \quad \text{for } 12 \le x \le 86 \tag{3.4}$$

(where k is a normalizing constant) as a linear approximation to Halley's table. However, as Haberman (1996) notes, a new era for the actuarial science started in 1825 with the law proposed by Benjamin Gompertz, the pioneer of a new approach to survival modelling. As it is well known, Gompertz's ideas can be properly expressed in terms of what we now call instantaneous intensity (or "force") of mortality. Denoting by μ_x the force of mortality, Gompertz's law is as follows:

$$\mu_x \;=\; \alpha\, e^{\beta\, x} \tag{3.5}$$

where α and β are positive parameters.

Gompertz's law constitutes one of the most influential proposals in the early times of survival modelling. Actually, many contributions in the field of mortality laws, throughout the latter half of the 19th century, generalize Gompertz's law or, anyhow, proceed from Gompertz's ideas. Remarkable examples are given by the laws proposed by William Makeham in 1860, Wilhelm Lazarus in 1867, Thorvald Thiele in 1867, Ludvig Oppermann in 1870.

Focussing on the problem of representing the mortality over the whole lifetime span, Thiele proposed the following function as the force of mortality:

$$\mu_x \;=\; \alpha_1\, e^{-\beta_1 x} + \alpha_2\, e^{-\beta_2 (x-\eta)^2} + \alpha_3\, e^{\beta_3 x} \tag{3.6}$$

where all parameters are positive (non-negative) real numbers. The first term on the right hand side of (3.6) represents the (decreasing) mortality at very young ages, the second term represents the mortality hump at young-adult ages, the third term (which coincides with Gompertz's law) represents the mortality at adult and old ages. With $\alpha_1 = \beta_1 = \beta_2 = 0$ we obtain Makeham's law, which can be written as follows:

$$\mu_x \;=\; \gamma + \alpha\, e^{\beta\, x} \tag{3.5$'$}$$

where $\alpha, \beta > 0$ and $\gamma \ge 0$. Clearly (3.5$'$) generalizes Gompertz's model. The two terms on the right side of (3.5$'$) can be interpreted as follows: $\alpha\, e^{\beta\, x}$ represents senescent mortality, whereas γ represents "background" mortality, which is by assumption independent of age.

Fitting formula (3.6) to experienced mortality is not a trivial matter (in particular at the time it was proposed). It is worth noting that, a century later, Heligman and Pollard (1980) proposed the same structure to model mortality odds:

$$\frac{q_x}{p_x} \;=\; A^{(x+B)^C} + D\, e^{-E(\ln x - \ln F)^2} + G\, H^x \tag{3.7}$$

In 1932, W.F. Perks proposed a family of survival models, represented by the following formula:

$$\mu_x = \frac{\alpha\ e^{\beta x} + \gamma}{\varepsilon\ e^{-\beta x} + \delta\ e^{\beta x} + 1} \tag{3.8}$$

Setting $\varepsilon = \delta = 0$, we find Makeham's law (3.5′), which is then a member of the family (3.8). Perk's models constitute the actuarial antecedent of heterogeneity modelling, as we will see in Section 3.4.3.

Functions other than the force of mortality have been also used in early survival modelling. For example, Wilhelm Lexis in 1877 represented the distribution of deaths by age (i.e. the "death curve") by a normal curve. Hence, he addressed the probability density function of the random lifetime.

A number of mathematical formulae used to represent the age pattern of mortality are reviewed by Benjamin and Pollard (1993).

3.2.2 Age-continuous models

Following the probabilistic structure laid down thanks to mathematical formulae fitting the experienced mortality, both actuarial theory and actuarial practice adopted an age-continuous approach to life insurance problems. In 1869, Wesley Woolhouse wrote the first complete presentation of life insurance mathematics on an age-continuous basis, considering sums assured payable at the moment of death as well as annuities payable continuously. One of the most remarkable contribution to age-continuous actuarial mathematics is the differential equation for the mathematical reserve proposed by Thorvald Thiele in 1875.

On the application side, it is worth noting, for instance, that at the beginning of the 20th century the Life office of Assicurazioni Generali in Trieste was equipped with a tariff system constructed on an age-continuous basis (see Graf, 1906). The underlying survival model was based on Makeham's law.

3.3 From single- to multiple-decrement and multistate models

When benefits only depend on the lifetime of the insured (or the annuitant), it is quite natural that the related actuarial model allows for just two "states", i.e. "alive" and "dead". Conversely, more states are needed, for example, to deal with:

- insurance products including benefits depending on the health status, e.g. disability benefits;
- death benefits whose amount depends on the cause of death, e.g. term assurance with a rider benefit in the case of accidental death;
- withdrawal benefits; etc.

As noted by Daw (1979), the work of Daniel Bernoulli, dated 1766, in which the effects of inoculation against smallpox are analyzed, constitutes a very important piece of actuarial pre-history, representing the starting point of multiple decrement modelling. Multiple decrement theory was progressed further in subsequent years. For example, the expression of the survival function $S(x)$ in terms of the forces of mortality, $\mu_t^{(i)}$, $i = 1, 2, \ldots, r$, related to r causes of death is attributable to A. Cournot and dates back to 1843:

$$S(x) \;=\; \exp\left(-\sum_{i=1}^{r}\int_0^x \mu_t^{(i)}dt\right) \tag{3.9}$$

Early actuarial applications of multiple decrement theory are related with the modelling of disability benefits. Notable contributions to the modelling of permanent disability annuities, throughout the second half of the 19th century, must be attributed to K.F. Heym, A. Wiegand, T. Wittstein, J. Karup (see Haberman, 1996; Seal, 1977). A systematic approach to disability benefits in both the continuous and discrete cases was presented in 1900 by E. Hamza, who also proposed a concise notation for disability-related functions, which has been widely adopted in the following decades.

Modelling non-necessarily permanent disability benefits requires a more complex probabilistic structure. This problem was attacked by R. Risser in 1912 and G. du Pasquier in 1912 and 1913. In particular, du Pasquier derived the differential equations for calculating transition probabilities from transition intensities. Hence, the proposed approach constitutes an early implementation of what is now denoted as the "transition intensity approach" (see Haberman and Pitacco, 1999). Moreover, du Pasquier's work presents an early application of Markov chains and lays the foundation for modern actuarial applications to disability insurance and other health-related insurance products.

Though the basic mathematics of what we now call a Markov chain model were developed during the 18th century (see Seal, 1977), an explicit and systematic use of the mathematics of Markovian multiple state models dates back to the end of the 1960s. The paper by Franckx (1963) constitutes a first step towards a unifying approach to the mathematics of the insurances of the person in terms of Markov chains. In Daboni (1964) a (age-discrete) Markov model is used to represent various types of annuities and the relevant random present values are analyzed.

The seminal contribution by Hoem (1969) places life and other contingencies within the framework of a general, unified probabilistic theory, using the Markov assumption. A time-continuous approach is adopted and formulae and theorems for actuarial values, premiums and reserves are derived. This contribution paves the way to age-continuous Markov modelling for the insurances of the person. See also Hoem (1988).

Transition intensities depending on the duration-in-state (i.e. inception-select intensities, in the actuarial language) can be represented in the framework of multiple state models adopting semi-Markov assumptions, as proposed by Hoem

(1972) in a time-continuous context. Conversely, Amsler (1968) represents duration-dependency in a time-discrete Markov context by splitting the states according to the duration of the presence; hence a semi-Markov structure is implemented in a Markov framework. Such a model is used, for instance, in the Dutch actuarial practice for representing the disability duration effect on the probabilities of recovery and death of disabled insureds.

Application of multiple state modelling to actuarial problems have been progressed further (for an extensive list of references see Haberman and Pitacco, 1999). At present multiple state modelling provides the actuary with a unifying approach to actuarial problems concerning a number of insurance covers within the area of the insurances of the person (life assurance and related rider benefits, annuities, disability covers, Long Term Care products, Critical Illness covers, etc.).

3.4 From homogeneity to heterogeneity

3.4.1 Some preliminary ideas

The awareness of heterogeneity in populations in respect of mortality can be traced back to the pre-history of survival modelling and actuarial science. As Hald (1987) notes, de Witt was aware that the nominee of an annuity is, with a high probability, a healthy person with a particularly low mortality in the first years of annuity payment and, generally, with an expected lifetime higher than the average (see also point (d) in Section 2.2). Thus, the presence of adverse selection (in particular due to "self-selection") must be recognized.

As far as explicit allowance for heterogeneity is concerned, the following aspects should be stressed. As Haberman (1996) notes, the earliest life tables developed for males and females were constructed (and published in 1740) by Nicholas Struyck, on the basis of the mortality among annuitants. The work of Francis Corbaux, dated 1833, constitutes a landmark towards heterogeneity. Corbaux dealt with a number of topics in the field of demography; among these, the idea that a population life table must be meant as a mixture of several life tables pertaining to various subgroups is particularly interesting, also from an actuarial point of view. Corbaux also singled out several variables (or "risk factors"), e.g. sex, occupation, etc., which sensibly influence the mortality of each individual and which should be allowed for in pricing insurance products, in order to avoid adverse selection.

Heterogeneity in a population of assured lives is also due to the effect of medical selection. When medical ascertainments lead to accept a risk at standard conditions, the insured's mortality is lower than the average, especially in the initial policy years. Thus, if q_x denotes the annual probability of death for a generic policyholder currently aged x, irrespective of the time elapsed since policy issue (the "aggregate" probability), whereas $q_{[x-t]+t}$ denotes (in the usual actuarial notation) the annual probability of death for a policyholder currently aged x with age $x-t$ at policy issue, the following inequalities are commonly assumed (as suggested by statistical evidence):

$$q_{[x]} < q_{[x-1]+1} < q_{[x-2]+2} < \dots \tag{3.10}$$

Probabilities $q_{[x-t]+t}$, $t = 0, 1, \dots$ are called "(issue-) select". Similar inequalities hold when self-selection among annuitants is concerned.

The presence of initial selection in an insured population was pointed out in 1850 by the English actuaries E. J. Farren and J. A. Higham, who also stressed that selection is likely to last for a limited time (5 to 10 years, say) only. Then, we can assume a selection period of t' years, whence

$$q_{[x]} < q_{[x-1]+1} < q_{[x-2]+2} < \dots < q_{[x-t']+t'} = q_{[x-t'-1]+t'+1} = \dots = q'_x \tag{3.11}$$

where q'_x denotes the common value of the probabilities $q_{[x-t]+t}$, $t \geq t'$. The q'_x's constitute the so-called "ultimate" table. Note that the aggregate probability q_x results in an average of the select probabilities $q_{[x-t]+t}$, $0 \leq t \leq t'-1$, and the ultimate probability q'_x. The combined use of select and ultimate tables was proposed by the English actuary T. B. Sprague in 1878.

The use of select (or ultimate-select) tables implies the dependence of actuarial values (e.g. premiums, reserves, etc.) on both the attained age and the duration since policy issue. However, in some actuarial calculations the effect of duration is much stronger than that of attained age. For example, this is the case of the mathematical reserve of an endowment insurance, for any given policy term (at least within a reasonable range of entry ages). So, Smolensky (1934) proposed the concept of "compact" table, a mortality table whose elements, $\bar{q}_t$, only depend on the policy duration t, to be used for evaluating endowment portfolio reserves. Clearly, the use of compact tables was particularly time saving compared with the use of tables involving select probabilities $q_{[x-t]+t}$. Improvements in computing power caused a lack of interest in compact tables. Notwithstanding, "compact schemes" are still used for actuarial calculations, e.g. concerning disability benefits, with the proviso that t must be meant as a duration since (disability) inception.

In more recent times, temporary initial selection and other durational effects have been dealt with in a multistate framework. The first contribution is given by Norberg (1988), in which a Markov model is defined in order to provide an explanation of the selection phenomenon. A more general model was then proposed by Møller (1990).

Turning to general aspects of heterogeneity in populations, we note that, in a modern perspective, heterogeneity should be approached addressing two main issues:

(i) detecting and modelling observable heterogeneity factors;

(ii) allowing for unobservable heterogeneity factors.

3.4.2 Observable heterogeneity factors

In the actuarial context, allowing for observable factors leads to risk classification, the "actuarial branch" of differential mortality. The earliest contributions to the analysis of mortality of impaired insured lives can be traced back to the last decades of the 19th century. In particular, in 1874 G. Humphreys reported about the mortality experience of a group of impaired lives accepted for life insurance at higher premium rates, with impairments classified into eight groups. The increase in premium rates was determined by assuming that the insured's age was higher than the real current age, hence adopting what is commonly called an "age shift".

In 1919, O. Rogers and A. Hunter of the New York Life Insurance Company proposed a new method for risk evaluation, called the "numerical rating system" (see Cummins et al, 1983). According to this method, each applicant is rated starting from an average rating, which is amended depending on the specific level of various risk factors: build (i.e. weight in relation to height), family record, occupation, habits, etc. Referring to annual probabilities of death, let q_x^{ave} represent the average probability. Denote by ρ_h the influence of the h-th risk factor, with $h = 1, 2, \ldots, m$; of course it may be $\rho_h \geq 0$ or $\rho_h < 0$. The specific rating is then given by

$$q_x^{spec} \; = \; q_x^{ave} \left(1 + \sum_{h=1}^{m} \rho_h \right) \tag{3.12}$$

Thus, the various risk factors are assumed to act additively in determining the total effect $\rho = \sum_{h=1}^{m} \rho_h$. This model probably represents the first sound contribution of medical statistics to actuarial practice. The numerical rating system is still used for underwriting in life insurance.

It is worth noting that, in the context of mortality laws, it is possible to find an interesting link between the rating structure whose result is represented by (3.12) and the age shifting mechanism proposed by Humphreys. Referring the rating structure to the force of mortality, let us assume

$$\mu_x^{spec} \; = \; \mu_x^{ave} \, (1 + \rho) \tag{3.13}$$

In the Gompertz hypothesis, for instance, we have (see (3.5)) $\mu_x^{ave} = \alpha \, e^{\beta \, x}$. Then

$$\mu_x^{spec} \; = \; \alpha \, e^{\beta \, x} \, (1 + \rho) \; = \; \alpha \, e^{\beta \, (x + \Delta x)} \; = \; \mu_{x+\Delta x}^{ave} \tag{3.14}$$

where Δx is implicitly defined by $e^{\beta \, \Delta x} = 1 + \rho$, and $\Delta x > 0$ if and only if $\rho > 0$. So, when impaired lives are concerned, the rating can be expressed by an age increment.

In very recent times, contributions to detecting and modelling heterogeneity factors have been provided by genetic testing. The interested reader should consult, for example, MacDonald (1999) and the very recent paper by Lemaire and MacDonald (2003) which provides an extensive review of the problems concerning application of genetic testing to insurance as well as an illustration of actuarial contributions to this area. As is well-known, the implementation of genetic testing as a tool for risk selection in life insurance is controversial.

3.4.3 Unobservable heterogeneity factors

It is worth noting that in the 1950's actuaries were aware of the consequences of heterogeneity due to unobservable factors, e.g. in terms of development of mortality over time in an insured population. The papers by Beard (1959) and Levinson (1959) probably constitute the earliest actuarial contributions to this topic.

Heterogeneity also affects the riskiness of a life insurance portfolio. Pollard (1970) deals with the assessment of riskiness in a portfolio which is assumed to consist of several groups with different mortality levels.

When allowing for unobservable heterogeneity factors, two approaches can be adopted:

(a) a discrete approach, which consists in expressing heterogeneity as a discrete mixture of terms; for example, see Levinson (1959) where an age-discrete approach is adopted with respect to mortality, Redington (1969) where an age-continuous approach to mortality is adopted;

(b) a continuous approach, according to which heterogeneity is expressed via a non-negative real valued variable, usually called "frailty" (or, in some early contributions, "longevity"); see Beard (1959, 1971), Vaupel, Manton and Stallard (1979); this approach has been recently adopted, in the actuarial context, by Butt and Haberman (2002).

Whatever the approach may be, the role of "frailty" is to include all unobservable factors influencing the individual mortality. What kind of information can be provided by modelling frailty and embedding it into a survival model? To understand the role of a frailty-based survival model, it is worth stressing the following aspects. The use of "laws" in survival modelling (see Section 3.2) aims at representing the effect of aging on the individual mortality. Different parameters (and even different laws) can be used to represent the age pattern of mortality for a "healthy" individual and, respectively, for impaired lives, with various impairment levels. Hence, if risk groups are built up on the basis of observable factors (health status, occupation, etc.), the mortality of individuals belonging to each group is represented by the appropriate mortality law. Notwithstanding, individuals inside each group may differ in respect of mortality, and this means that a residual heterogeneity affects the groups. A fortiori, a higher degree of heterogenity affects populations for which no splitting into risk groups has been worked out. Hence, in both cases information about the "average" age pattern of mortality (inside the group, or inside the population) may be of interest. Frailty-based survival models can produce this information.

In what follows, we only give a concise illustration of the continuous approach to frailty, following the pioneering ideas of Beard (1959), who should be considered a forerunner in mortality modelling allowing for frailty (see also Beard, 1971), the seminal contribution by Vaupel, Manton and Stallard (1979) and the presentation by Butt and Haberman (2002). For a general approach to heterogeneity models in the actuarial field the reader can consult Cummins et al (1983).

Referring to a generic individual in a given (heterogeneous) cohort, let us assume that his/her frailty remains constant throughout the whole life span. Let us consider the cohort at age x, $x \geq 0$, and denote by Z_x the random frailty (a positive real number) of an individual in the cohort. Note that, regardless the assumption of constant individual frailty, the distribution of frailty in the cohort does depend on the age, and this justifies the suffix x to denote the random frailty.

Let $\mu_x(z)$ denote the force of mortality conditional on the frailty level z, i.e.

$$\mu_x(z) = \lim_{t \to 0} \frac{\Pr\{T_x \leq t | Z_x = z\}}{t} \tag{3.15}$$

Assume then

$$\mu_x(z) = z\ \mu_x \tag{3.16}$$

where μ_x denotes the "standard" force of mortality, i.e. the force of mortality of a "standard" individual, conventionally with a (constant) unitary level of frailty. Assumption (3.16) expresses the so-called "multiplicative model". Some important results within the multiplicative approach to frailty follow.

Let $g_x(z)$ denote the probability density function (pdf) of the random frailty in the cohort at age x, Z_x. Given the "initial" pdf $g_0(z)$ and the standard force of mortality μ_x, it is possible to derive the pdf $g_x(z)$ for all x (see, for example, Butt and Haberman (2002)). Then, define the expected frailty in the cohort at age x:

$$\bar{z}_x \ = \ E(Z_x) \ = \ \int_0^{+\infty} z\ g_x(z)\ dz \tag{3.17}$$

Further, let $\bar{\mu}_x$ denote the "cohort force of mortality", namely the average force of mortality in a cohort whose members are aged x. It is possible to prove that:

$$\bar{\mu}_x \ = \ \mu_x\ \bar{z}_x \tag{3.18}$$

and that

(1) $\bar{z}_x$ decreases as x increases;

(2) $\bar{\mu}_x$ increases (as x increases) less rapidly than μ_x.

An intuitive interpretation of result (1) is that the average frailty decreases because people with a high frailty level have a high probability of dying early. This means, as indicated by result (2), that the "average" force of mortality increases, as the age increases, less rapidly than the standard individual force of mortality. As noted by Vaupel, Manton and Stallard (1979), this implies that studies concerning the human ageing may be biased when based on cohort survival data. In particular, the individual expected lifetime may be overestimated if the relevant evaluation relies on cohort experience.

To define a frailty-based survival model it is necessary to choose:

(a) the survival model for standard individuals, i.e. the force of mortality μ_x;
(b) the pdf $g_0(z)$.

As pointed out by Butt and Haberman (2002), usual choices for (a) are the Gompertz and Makeham laws, and for (b) the gamma and the inverse Gaussian distribution. It is worth noting that the combined choice of the Gompertz or the Makeham law and the gamma distribution leads to a cohort force of mortality $\bar{\mu}_x$ belonging to the family of survival models proposed by W. F. Perks in 1932 (see formula (3.8)), as proved by Beard (1959).

In Butt and Haberman (2002) an insurance application of frailty-based survival models is proposed. In particular, the authors discuss various choices and fit some models to two sets of life insurance mortality data. The obtained results suggest the use of the Gompertz - gamma model.

In Horiuchi and Wilmoth (1998) frailty models are used to explain the deceleration in the age pattern of mortality at very old ages. In particular, the authors explore the heterogeneity hypothesis by testing empirically some of its predictions.

4. MODELLING DYNAMIC MORTALITY

4.1 Mortality trends

Mortality experience over the last decades shows some aspects affecting the shape of curves representing the mortality as a function of the attained age. In particular (see Olivieri, 2001):

(a) an increasing concentration of deaths around the mode (at old ages) of the curve of deaths (the graph of the probability density function of the random lifetime, in an age-continuous setting) is evident; so the graph of the survival function moves towards a rectangular shape, whence the term "rectangularization" to denote this aspect;
(b) the mode of the curve of deaths (which, owing to the rectangularization, tends to coincide with the maximum age ω) moves towards very old ages; this aspect is called "expansion" of the survival function;
(c) higher levels and a larger dispersion of accidental deaths at young ages (the so-called young mortality hump) have been more recently observed.

Further aspects of mortality trends can be captured looking at the behaviour, for each integer age x, of the annual probability of death q_x drawn from a sequence of life tables pertaining to the same kind of population (e.g. males living in a given country). The graph constructed plotting the q_x's against time is usually called "mortality profile". Mortality profiles are often decreasing, in particular at adult and old ages.

Rectangularization and expansion phenomena and decreasing mortality profiles are witnessed by mortality experience in a number of countries. The reader can refer to MacDonald et al (1998) and Rüttermann (1999) for interesting international comparisons.

4.2 Mortality in a dynamic context: some basic ideas

Mortality improvements could induce underestimation of liabilities related to life annuities and other living benefits. So, trends in mortality imply the use of "projected" survival models for several actuarial purposes, e.g. for pricing and reserving as well as for assessing solvency in life offices and pension plans. A projected survival model aims at describing future age patterns of mortality, on the basis of the experienced mortality trend.

A dynamic approach to mortality underpins projected survival models. When working in a dynamic context (in particular when projecting mortality), the basic idea is to express mortality as a function of the (future) calendar year y. When a single-figure representation of mortality is concerned, a dynamic model is a real-valued function $\Psi(y)$. For example, the expected lifetime for a newborn, denoted by $\overset{\circ}{e}_0$ in a non-dynamic context, is represented by $\overset{\circ}{e}_0(y)$, a function of the calendar year y (namely the year of birth), when the mortality trend is allowed for. Similarly, the general death rate in a given population can be represented by a function $q(y)$, where y denotes the calendar year in which the population is considered.

In actuarial calculations, age-specific measures of mortality are usually needed. Then in a dynamic context, mortality is assumed to be a function of both the age x and the calendar year y. In a rather general setting, a dynamic survival model is a function $\Gamma(x, y)$, usually with real values. However, a vector-valued function is concerned if, for example, causes of death are allowed for.

In concrete terms, a real-valued function $\Gamma(x, y)$ may represent mortality rates, mortality odds, a force of mortality, a survival function, some transform of the survival function, etc. The projected survival model is given by the restriction $\Gamma(x, y)|y > y'$, where y' denotes the current calendar year, or possibly the year for which the latest (reliable) period life table is available. The projected survival model is constructed (and, in particular, the relevant parameters are estimated) by applying appropriate statistical procedures to past mortality experience.

Although age-specific functions are needed in actuarial calculations, the interest of single-figure indexes as functions of the calendar year should not be underestimated. In particular, important features of past mortality trends can be singled out focussing on the behaviour of some indexes meant as "markers" of the probability distribution of the random lifetime at birth, T_0 (or at some given age x, T_x). Examples of markers providing a "location" measure, and hence information about expansion, are as follows (notation refers to a non-dynamic context):

(1) the expected lifetime for a newborn, $\overset{\circ}{e}_0$;

(2) the expected lifetime at some fixed age x_0, $\overset{\circ}{e}_{x_0}$;

(3) the mode (at adult ages) of the curve of deaths, also called Lexis point.

Variability measures, from which information about the rectangularization process can be obtained, are given by:

(4) the variance of the random lifetime, $Var(T_0)$, or its standard deviation, $\sigma_0 = \sqrt{Var(T_0)}$;

(5) the interquartile range, $IQR = x'' - x'$, where x' and x'' are ages such that $S(x') = 0.75$ and $S(x'') = 0.25$, $S(x)$ denoting the survival function; note that IQR decreases as the lifetime distribution becomes less dispersed;

(6) the index H defined as follows: $H = -\frac{\int_0^\omega S(x)\, \ln S(x)\, dx}{\int_0^\omega S(x)\, dx}$; thus, H is minus the mean value of $\ln S(x)$, weighted by $S(x)$; as deaths become more concentrated, the value of H declines and, in particular, $H = 0$ if the survival function has a perfectly rectangular shape.

For a deep analysis about the role of several indexes in representing the rectangularization process, the reader should refer to Wilmoth and Horiuchi (1999). For more information about the index H, sometimes referred to as the "entropy" of the survival function, see Keyfitz (1985). Further information about the probability distribution of the random lifetime is provided, for instance, by:

(7) the probability for a newborn of dying before a given age x_1, ${}_{x_1}q_0 = 1 - S(x_1)$; for x_1 small (say 1, or 5), a measure of infant mortality is then obtained;

(8) the so-called "endurance", i.e. the age ξ such that $S(\xi) = 0.90$.

In a dynamic context, all markers should be noted addressing the calendar year y, e.g. $\sigma_0(y)$, $IQR(y)$, $H(y)$, $\xi(y)$, etc.

Turning back to age-specific functions, assume now that both age x and calendar year y are discrete variables. Hence, $\Gamma(x, y)$ can be represented by a matrix whose rows correspond to ages and columns to calendar years. For example, let $\Gamma(x, y) = q_x(y)$. Then, the annual probabilities of death in the matrix can be read according to three arrangements:

(a) a "vertical" arrangement (i.e. by columns),

$$q_0(y),\ q_1(y), \ldots,\ q_x(y), \ldots \tag{4.1}$$

corresponding to a sequence of period life tables, each table referring to a given calendar year y;

(b) a “diagonal” arrangement,

$$q_0(y),\ q_1(y+1), \ldots,\ q_x(y+x), \ldots \tag{4.2}$$

corresponding to a sequence of cohort life tables, each table referring to the cohort born in year y;

(c) a “horizontal” arrangement (i.e. by rows),

$$\ldots, q_x(y-1),\ q_x(y),\ q_x(y+1), \ldots \tag{4.3}$$

yielding the mortality profiles, each profile referring to a given age x.

As will emerge from the discussion of some contributions, thinking in terms of the various arrangements can also help in understanding different approaches to the interpolation of mortality data.

4.3 Mortality projections: from the forerunners to research in progress

4.3.1 The antecedents

As noted by Cramér and Wold (1935), the earliest attempt to project mortality is probably due to the Swedish astronomer H. Gyldén. In a work presented to the Swedish Assurance Association in 1875, he fitted a straight line to the sequence of general death rates of the Swedish population concerning the years 1750 to 1870. A similar graphical interpolation was proposed in 1901 by T. Richardt for sequences of the annuity values a_{60} and a_{65}, calculated according to various Norwegian life tables, and then projected via extrapolation for application to pension plan calculations. Note that, as in the proposal by Gyldén, also in this case the projection of a single-figure index was concerned.

Mortality trends and the relevant effects on life assurance and pension annuities were clearly perceived at the beginning of the 20th century, as witnessed by various initiatives in the actuarial field. In particular, it is worth noting that the subject “Mortality tables for annuitants” was one of the topics discussed at the 5th International Congress of Actuaries, held in Berlin in 1906. Nordenmark (1906), for instance, pointed out that improvements in mortality must be carefully considered when pricing life annuities and, in particular, cohort mortality should be addressed to avoid underestimation of the related liabilities. The 7th International Congress of Actuaries, held in Amsterdam in 1912, included the subject “The course, since 1800, of the mortality of assured persons”.

As Cramér and Wold (1935) notes, a life table for annuities was constructed in 1912 by A. Lindstedt, who used data from Swedish population experience and, for each age x, extrapolated the sequence of annual probability of death, namely the mortality profile $q_x(y)$, hence adopting a “horizontal” approach. Probably, this work constitutes the earliest projection of age-specific functions.

4.3.2 Early seminal contributions

Blaschke (1923) proposed a Makeham-based projected survival model. A dynamic Makeham's law was defined as follows:

$$\mu_x(y) \;=\; \gamma(y) + \alpha(y)\; \beta(y)^x \tag{4.4}$$

Hence, the three parameters are functions of the calendar year y. For the projection, a "vertical" method was proposed, consisting in the estimation of the parameters for each period table (or "cross sectional" table) based on the experienced mortality, and then in fitting the estimated values; projected values of the three parameters are obtained via extrapolation. Note that when a law-based projected survival model is used, as in (4.4), the age pattern of mortality depends on the calendar year y via the parameters of the mathematical law only.

As Cramér and Wold (1935) notes, in 1924 the Institute of Actuaries in London proposed a "horizontal" method for mortality projection, assuming for the annual probability of death the following expression:

$$q_x(y) \;=\; a_x + b_x\; c_x^y \tag{4.5}$$

thus, $q_x(y)$ is an exponential function of the calendar year y, whence the name "exponential formula" commonly used to denote this approach to mortality projections. Parameters a_x, b_x and c_x are estimated on the basis of observed mortality profiles.

It is worth noting that projection formulae currently used by UK actuaries for annuitants and pensioners mortality tables are particular cases of formula (4.5). For instance, with $a_x = 0,\; b_x = q_x(y')\, r_x^{-y'},\; c_x = r_x$, where y' denotes the current year and r_x represents the annual rate of mortality improvement (if $r_x < 1$) at age x, the so-called "reduction factor", we obtain

$$q_x(y) \;=\; q_x(y')\; r_x^{y-y'} \tag{4.5a}$$

Moreover, with $a_x = \lambda_x\, q_x(y'),\; b_x = (1-\lambda_x)\, q_x(y')\, r^{-y'},\; c_x = r$, we find

$$q_x(y) \;=\; q_x(y') \left[\lambda_x + (1-\lambda_x)\; r^{y-y'}\right] \tag{4.5b}$$

where $\lambda_x\, q_x(y')$ represents (if $r < 1$) the asymptotic mortality at age x; in this case the speed of convergence, and hence r, is assumed to be independent of age. The reader should refer to CMIR10 (1990) and CMIR17 (1999) for more details. The formula proposed in 1929 by the German actuary C. W. Sachs also represents a particular case of (4.5), being as follows:

$$q_x(y) \;=\; q_x(y')\; a^{\frac{y-y'}{x+b}} \tag{4.5c}$$

where a and b are constants.

Let us turn to the "diagonal" approach. In 1927 A. R. Davidson and A. R. Reid proposed a Makeham-based model, with a dynamic Makeham's law defined as follows:

$$\mu_x(y) \;=\; \delta(\tau) + \varphi(\tau)\;\psi(\tau)^x \tag{4.6}$$

where $\tau = y - x$ denotes the year of birth. In the implementation, $\psi(\tau) = \psi$ was assumed for all τ, whereas the functions $\delta(\tau)$ and $\varphi(\tau)$ were estimated via a cohort graduation (see Davidson and Reid, 1927).

The use of Makeham-based projected survival models is discussed by Cramér and Wold (1935), dealing with graduation and extrapolation of Swedish mortality. In particular, the diagonal and the vertical approach are compared.

The assumption formulated in 1934 by Kermack, McKendrick and McKinlay constitutes another example of the diagonal approach to mortality projections. As Pollard (1949) notes, these authors showed that, for some countries, it was reasonable to assume that the force of mortality depended on the attained age x and the year of birth τ, and they deduced that

$$\mu_x(y) \;=\; Q(x)\,R(\tau) \tag{4.7}$$

where $Q(x)$ is a function of age only and $R(\tau)$ is a function of year of birth only.

Turning back to law-based projected models, it should be stressed that functions other than the force of mortality can be addressed. For example, Beard (1952) built up a projected model fitting a Pearson Type III curve to the curve of deaths and then taking some parameters (in particular the maximum age) as functions of the year of birth.

4.3.4 Some modern contributions

Seminal contributions to survival modelling and mortality projections have been produced by demographers, throughout the second half of the 20th century. The "optimal" table, model tables and relational methods probably constitute three of the most influential proposals in recent times, in the framework of survival analysis.

From previous sections, it clearly emerges that a number of projection methods are based on the extrapolation of observed mortality trends. Important examples are provided by formulae (4.4), (4.5) and (4.6). Albeit it seems quite natural that mortality forecasts are based on past mortality observations, different approaches to the construction of projected tables can be adopted.

Let us suppose that the existence of an "optimal" life table is assumed. The relevant age pattern of mortality must be meant as the limit to which mortality improvements can lead. Let q_x^* denote the limit probability of death at age x, whereas $q_x(y')$ denotes the current mortality. Assume then that the projected mortality $q_x(y)$ is expressed as follows:

$$q_x(y) = F[q_x^*, q_x(y')] \tag{4.8}$$

where the symbol F denotes some interpolation formula. In particular, an exponential interpolation can be adopted, leading for example to:

$$q_x(y) = q_x^* + (q_x(y') - q_x^*)\, r^{y-y'} \tag{4.8$'$}$$

with $r < 1$. Note that formula (4.5b) can be easily linked to (4.8′), choosing λ_x such that $q_x(y')\lambda_x = q_x^*$.

The idea of an "optimal" table was proposed by Bourgeois-Pichat (1952). The question was: "can mortality decline indefinitely or is there a limit, and if so, what is this limit?" Determining a limit table requires a number of assumptions about the trend in various mortality causes, so that an analysis of mortality by causes of death is required.

When a mortality law is used to fit observed data, the age pattern of mortality is summarized by some parameters (two or three, for Gompertz's law and Makeham's law respectively); see, for example, (4.4) and (4.6). Then, the projection procedure can be applied to the set of parameters (instead of the set of age-specific mortality rates), with a dramatic reduction in the "dimension" of the forecasting problem, namely in the number of "degrees of freedom". However, the age pattern of mortality can be summarized without resorting to mathematical laws (and hence avoiding the relevant choice). In particular, some typical values, or "markers", of the mortality pattern can be used to this purpose, as mentioned in Section 4.2.

The possibility of summarizing the age pattern of mortality by using some markers underpins the use of "model tables" in mortality projections. The first set of model tables was constructed in 1955 by the United Nations. The set was indexed on the expectation of life at birth, $\mathring{e}_0$, so that each table was summarized by the relevant value of this marker.

Model tables can be used for mortality forecasts as follows. A set of model tables is chosen, representing the mortality in a given population at several epochs, and assumed to represent also future mortality for that population. Trends in some markers are analyzed and then projected, possibly using some mathematical formula, to predict their future values. Projected age-specific mortality rates are then obtained entering the system of model life tables for the various projected values of the markers.

A new way to mortality forecasts was paved by the "relational method" proposed by W. Brass (see Brass, 1974), who focussed on the logit transform of the survival function, namely

$$\Lambda_x = \frac{1}{2} \ln\left(\frac{1 - S(x)}{S(x)}\right) \tag{4.9}$$

Brass noted empirically that Λ_x can be expressed in terms of the logit pertaining to a "standard" population, Λ_x^{stand}, via a linear relation, i.e.

$$\Lambda_x = \alpha + \beta\, \Lambda_x^{stand} \tag{4.10}$$

whose parameters are (almost) independent of age.

For the purpose of forecasting mortality, equation (4.10) can be used in a dynamic sense. In a dynamic survival modelling context, the Brass logit transformation is particularly interesting when applied to cohort data, as the logits referring to successive birth-year cohorts seem to be linearly related (see Pollard, 1987). Hence, denoting by $\Lambda_x(\tau)$ the logit of the survival function for the cohort born in the calendar year τ, $S(x,\tau)$, we have:

$$\Lambda_x(\tau) = \frac{1}{2} \ln\left(\frac{1 - S(x,\tau)}{S(x,\tau)}\right) \tag{4.9$'$}$$

Referring to a couple of birth years, τ_k and τ_{k+1}, assume

$$\Lambda_x(\tau_{k+1}) = \alpha_k + \beta_k\, \Lambda_x(\tau_k) \tag{4.10$'$}$$

So, the problem of projecting mortality reduces to the problem of extrapolating the two series α_k and β_k. Projected values of various life table functions can be derived from the inverse logit transformation:

$$S(x,\tau) = \frac{1}{1 + \exp[2\, \Lambda_x(\tau)]} \tag{4.9$''$}$$

A different transform of the survival function $S(x)$ has been addressed by L. Petrioli and M. Berti. The proposed transform is the "resistance function" (see Petrioli and Berti, 1979; Keyfitz, 1982), defined as follows:

$$r(x) = \frac{\frac{S(x)}{\omega - x}}{\frac{1 - S(x)}{x}} \tag{4.11}$$

where ω denotes the maximum age. Thus, the transform is the ratio of the average annual probability of death beyond age x to the average annual probability of death prior to age x. The resistance function has been graduated with the curve:

$$r(x) = x^{\alpha}(\omega - x)^{\beta} e^{Ax^2 + Bx + C} \tag{4.12}$$

and, in particular, with the three-parameter curve:

$$r(x) = k\, x^{\alpha}(\omega - x)^{\beta} \tag{4.12$'$}$$

Model tables have been constructed on combinations of the three parameters, focussing on the values of some markers.

In a dynamic context, the mortality trend is represented assuming that (some of) the parameters of the resistance function depend on the calendar year y. Experienced mortality trends lead to parameters fitting through time, so that, referring to equation (4.12′), we have:

$$r(x,y) \;=\; k(y)\; x^{\alpha(y)}(\omega - x)^{\beta(y)} \tag{4.13}$$

Note that, assuming a model for the resistance function (see (4.12) and (4.12′)), the resulting projection model can be classified as an analytical model, even though it does not directly address the survival function.

The Petrioli-Berti model has been used to project the mortality of the Italian population, and then has been adopted by the Italian Association of Insurers to build up projected mortality tables for annuity business.

In the last decades of the 1900's, various mortality law-based projection models have been proposed. Forfar and Smith (1988) have performed mortality projections using the Heligman-Pollard law, assuming that various relevant parameters are functions of the calendar year (see also Benjamin and Soliman, 1993). Poulin (1980) has proposed a Makeham-based projection formula, whereas Wetterstrand (1981) has used Gompertz's law.

Modelling of mortality reduction factors (see Section 4.3.2) has been recently dealt with by Renshaw and Haberman (2000) and Sithole et al (2000).

4.3.5 From present contributions to research in progress

In the 1990's, a new method for forecasting the age pattern of mortality was proposed and then extended by L. Carter and R.D. Lee (see Lee and Carter, 1992; Lee, 2000). The Lee-Carter (LC) method addresses the central death rate, to represent the age-specific mortality. Let $m_x(y)$ be the central death rate for age x at time y. The model is as follows:

$$\ln m_x(y) \;=\; a_x + b_x\, k_y + e_{x,y} \tag{4.14}$$

where the a_x's describe the age pattern of mortality averaged over time, whereas the b_x's describe the deviations from the averaged pattern when the coefficient k_y varies. The variation in the level of mortality with y is described by k_y. Finally, the quantity $e_{x,y}$ denotes the error term.

Parameters a_x, b_x and k_y are estimated from experienced mortality, obtaining the estimates $\hat{a}_x$, $\hat{b}_x$, $\hat{k}_y$ (see also Renshaw and Haberman, 2003b). Forecasts follow by modelling the values of k_y as a time series, e.g. a random walk with drift. Starting from a given year y', forecasts of mortality rates are then computed, for $s = 1, 2, \dots$, as follows:

$$m_x(y'+s) \;=\; \exp(\hat{a}_x + \hat{b}_x\, k_{y'+s}) \;=\; m_x(y')\;\exp\left[\hat{b}_x(k_{y'+s} - \hat{k}_{y'})\right] \tag{4.15}$$

An important feature of the LC methodology should be stressed. Traditional projections models provide the forecaster with point estimates of future mortality rates (or other age-specific quantities). On the contrary, the LC method allows for uncertainty in forecasts. In fact, $m_x(y)$ is modelled as a stochastic process driven by the stochastic process k_y, whence interval estimates can be computed for the projected values of mortality rates.

The LC methodology represents one of the most influential proposals in recent times, in the field of mortality forecasts. In very recent times, indeed, much research work as well as many applications to actuarial problems are directly related to this methodology. See the list of references in Lee (2000); see also Renshaw and Haberman (2003a, 2003b), Brouhns and Denuit (2002).

4.3.6 Some remarks

Methods for mortality projections can be classified according to various points of view. For brevity, we only focus on two criteria. Whatever the approach may be, mortality forecasts are obviously based on observed data, which usually consist in (cross-sectional) mortality tables. As regards the "use" of data in extrapolating observed trends, the following classification seems to be interesting.

(1) Age-specific data can be directly used for mortality forecasts. Thus, the projection procedure is applied to quantities such as mortality rates q_x (see, for example, the exponential formula in Section 4.3.2), central mortality rates m_x (see the LC methodology), mortality odds $\frac{q_x}{p_x}$, etc.

(2) Data can be "summarized" in several ways. Important examples are provided by the use of mortality laws (for instance Makeham's law; see Section 4.3.2), by model tables and by the Brass relational method. In these cases, the projection procedure is applied to the parameters of the law, or the markers associated to the model tables, or the parameters of the Brass linear relation.

Advantages and disadvantages of the two approaches constitute a controversial matter. For interesting discussions on this issue, the reader should consult Keyfitz (1982) and Pollard (1987).

As far as the link between experience data (i.e. mortality tables) and projected mortality is concerned, it is worth noting that:

- mortality tables provide estimates of random mortality in a (past) population;
- mortality in a future population is random, also because of its unknown trend.

The stochastic nature of mortality should not be disregarded, in particular when forecasts are concerned. As regards the allowance for stochastic mortality, we can note what follows.

(a) Traditional projection methods disregard the stochastic nature of (observed) mortality and provide the forecaster only with "point" estimates of future mortality.

(b) The stochastic model underlying the LC methodology recognizes the observed mortality rates as estimates, and allows for interval estimation of future mortality rates.

(c) Uncertainty in future mortality rates is first attributable to random fluctuations around the relevant point estimates, namely to "process risk". Moreover, deviations may be attributed also to the choice of the projection model, because the relevant parameters or the structure of the model itself do not reflect the actual mortality trend. Hence, a "parameter risk" and a "model risk" should be allowed for when projecting mortality. This topic is dealt with in Section 5.

4.3.7 Further issues of mortality forecasts

We just mention some issues of mortality forecasts which are beyond the scope of this article. The projection methods we have described refer to mortality in aggregate. Notwithstanding, many of them can be used to project mortality from different causes separately. A cause-of-death projection study was proposed by A.H. Pollard in 1949, based on Australian population data (see Pollard, 1949). Projections by cause of death offer a useful insight into the changing incidence of the various causes. Conversely, some important problems arise when this type of projection is adopted. In particular, it should be stressed that complex interrelationships exist among causes of death, whilst the assumption of independence is commonly accepted. For example, heart diseases and lung cancer are correlated, as both are linked to smoking habits. A further problem concerns the difficult identification of the cause of death for elderly people. For these reasons, many forecasters prefer to carry out mortality projections only in aggregate terms.

Projecting mortality of very old population segments deserves special attention. A first problem obviously concerns the observed old-age mortality rates, heavily affected by random fluctuations because of their scarcity. Analytical models provide a smoothing tool, but they may fail in representing the very old-age mortality. In particular, it has been observed that the exponential rate of mortality increase at very old ages is not constant, as in Gompertz's law, but declines. Thus, shifting from the Gompertz or Makeham assumptions may be necessary in order to fit the old age pattern of mortality. For this topic the reader can refer to Horiuchi and Wilmoth (1998), where the problem is attacked in the context of the frailty models (see also Section 3.4.3). Of course, these problems also affect projection procedures, in both an age-specific and a mortality law-based approach. For more information the reader should consult, for example, Buettner (2002) and the bibliographic references therein.

When projecting mortality, collateral information available to the forecaster can be allowed for. Information may concern, for example, trends in smoking habits, trends in prevalence of some illness, improvements in medical knowledge and surgery, etc. Thus, projections can be performed according to an assumed scenario. Obviously, some degree of arbitrariness follows, affecting the results.

The reader interested in various perspectives on forecasting mortality can refer to Tabeau et al (2001), in which a number of approaches to mortality projections are discussed and several applications are described.

5. THE LONGEVITY RISK

5.1 Modelling the uncertainty in mortality trends

As discussed in Section 4, observations on mortality suggest to adopt projected mortality tables for the actuarial appraisal of annuities and other living benefits, i.e. to use demographic hypotheses including a forecast of future mortality trends. Notwithstanding, whatever hypothesis is assumed the future trend is random, whence longevity risk arises. Then, the probabilistic model used for representing mortality should allow for the assessment of longevity risk. This can be obtained as described below.

Following the notation introduced in Section 4, let $\Gamma(x, y)$ denote a projected mortality model, i.e. a (real-valued) function of age x and calendar year y, which expresses, in some way, the mortality of people aged x in the (future) calendar year y, i.e. born in year $\tau = y - x$. A family of projected mortality models can express several alternative hypotheses about future mortality evolution. To this purpose, denote by $K(\tau)$ a given assumption about the mortality trend for people born in year τ, and by $\mathcal{K}(\tau)$ the set of such hypotheses. Then, the family of projected models

$$\{\Gamma[x, \tau + x | K(\tau)];\ K(\tau) \in \mathcal{K}(\tau)\} \tag{5.1}$$

should be addressed, with $\Gamma[x, \tau + x|K(\tau)]$ representing the projected model conditional on the specific hypothesis $K(\tau)$.

If the mortality is described through a mathematical law, such as those proposed by Gompertz, Makeham, Thiele, Weibull, etc., the law itself is characterized by a vector-valued parameter $\theta(\tau)$ (e.g. a couple, as in the Gompertz and Weibull laws). An appropriate choice of the vector-valued function $\theta(\tau)$ may reflect a given hypothesis about the mortality trend. In a parametric context, since the age pattern of mortality depends on time only via the parameter, the family (5.1) can be simply denoted as follows:

$$\{\Gamma[x \,|\, \theta(\tau)];\ \theta(\tau) \in \Theta(\tau)\} \tag{5.2}$$

where $\Theta(\tau)$ denotes the parameter space and, as before, $\Gamma[x \mid \theta(\tau)]$ is the projected model conditional on the mortality trend hypothesis expressed by the function $\theta(\tau)$.

In the following, for simplicity we refer to one cohort only, i.e. a given year of birth. So, τ can be dropped from the notation. Therefore, the family of models is simply

$$\{\Gamma[x|K];\ K \in \mathcal{K}\} \tag{5.1$'$}$$

and in the parametric context

$$\{\Gamma[x\,|\,\theta];\ \theta \in \Theta\} \tag{5.2$'$}$$

It is understood that spaces $\mathcal{K}$ and, respectively, Θ refer to the specific cohort addressed.

The parameter space can be either a discrete or a continuous set. In the former case, referring in particular to a finite set we have

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_m\} \tag{5.3}$$

where m choices are made for the parameter, each one expressing a particular mortality trend. Trivially, if $m = 1$ we are dealing with one (projected) model; hence, uncertainty in future mortality is not allowed for. Conversely, if $m > 1$ we are dealing with several projected models, expressing alternative views on future mortality trends. According to the elements of Θ, one can perform a scenario testing, assessing the range of variation of quantities such as profits, reserves, etc. Hence, the sensitivity of such quantities to a change in mortality evolution is investigated.

As an alternative, assign a weight g_j to each parameter value θ_j in Θ (i.e. to each projection hypothesis), such that $\sum_{j=1}^{m} g_j = 1$. Hence, the set $\{g_j\}_{j=1,2,\dots,m}$ can be interpreted as a probability distribution assigned on Θ.

So a stochastic approach can be adopted, calculating unconditional variances, percentiles, etc., of the value of future benefits. Whilst in the previous approach only the risk of random fluctuations is accounted for, in this latter case both the risk of random fluctuations and the longevity risk are considered. The relevant actuarial evidence is discussed in Section 5.3.

In order to have a better understanding of the modelling structure described above, let us consider the residual lifetime T_x. According to a given choice θ_j of the parameter, the (conditional) moments of T_x may be evaluated. In an age-continuous context, we have for example:

$$E(T_x|\theta_j) = \int_0^{+\infty} t\ f_x(t|\theta_j)\ dt \tag{5.4}$$

$$Var(T_x|\theta_j) = \int_0^{+\infty} (t - E(T_x|\theta_j))^2\ f_x(t|\theta_j)\ dt \tag{5.5}$$

where $f_x(t|\theta_j)$ is the (conditional) pdf of the residual lifetime T_x.

The unconditional moments are then as follows:

$$E(T_x) = \int_0^{+\infty} \sum_{j=1}^{m} t \; f_x(t|\theta_j) \; g_j \; dt \tag{5.6}$$

$$Var(T_x) = \int_0^{+\infty} \sum_{j=1}^{m} (t - E(T_x))^2 \; f_x(t|\theta_j) \; g_j \; dt \tag{5.7}$$

In particular, developing expression (5.7) the following well-known result can be obtained for the variance:

$$Var(T_x) = \sum_{j=1}^{m} Var(T_x|\theta_j) \; g_j + \sum_{j=1}^{m} (E(T_x|\theta_j) - E(T_x))^2 \; g_j \tag{5.8}$$

or also

$$Var(T_x) = E(Var(T_x|\tilde{\theta})) + Var(E(T_x|\tilde{\theta})) \tag{5.8$'$}$$

where $\tilde{\theta}$ denotes the random value of the parameter. Note in particular that whilst (5.5) and the first term in (5.8) (or (5.8$'$)) measure the risk of random fluctuations, the second term in (5.8) (or (5.8$'$)) measures the risk of systematic deviations, i.e. the longevity risk.

The parameter space Θ can be a continuous set, i.e.

$$\Theta = A_1 \times A_2 \times \ldots \times A_n \tag{5.9}$$

where $A_1, A_2 \ldots, A_n$ denote the intervals of possible values for each of the n items of the parameter characterizing the mortality law we are dealing with (i.e. for each element in the vector-valued parameter θ). A pdf $g(\theta)$ can be assigned on Θ, with $\int_\Theta g(\theta) \; d\theta = 1$, in order to perform unconditional evaluations. Results similar to (5.4)–(5.8$'$) obviously hold.

It is worth noting that, although we assume the (usual) hypothesis of stochastic independence among the individual lifetimes T_x conditional on any given K (or θ), the lifetimes are unconditionally correlated.

5.2 Process risk, parameter risk, model risk

In Section 4.3.6, while addressing the stochastic nature of (past and future) mortality, we already mentioned the types of risk related to mortality deviations from expected values. We now turn back to this issue, which must be carefully considered when dealing with the riskiness of an insurance portfolio (or a pension plan), namely with the actuarial evidence of uncertainty in future mortality.

In a life insurance portfolio (or a pension plan), deviations from expected mortality in future years may obviously occur. How to explain deviations from

expected frequencies? Assume that expected frequencies are calculated from a projected life table. The following items are then involved in calculations:
- a projection model, e.g. an exponential formula with reducing factors r_x (see Section 4.3.2);
- a parameter in the projection model, viz. the set of r_x's;
- a survival model for each (future) year y, namely a life table (the y-entry of the projected life table).

Deviations from expected values arise from the following risk sources (see Cairns, 2000):

(a) the stochastic nature of a given model (the number of deaths, in each calendar year y, is a random variable); thus, the so-called "process risk" arises;
(b) uncertainty in the values of the parameters (reduction factors do not reflect the actual trend), originating the "parameter risk";
(c) uncertainty in the model underlying what we can observe (actual trend is not represented by an exponential decline in mortality rates); hence, the "model risk" arises.

So, we can recognize

(1) a random component arising from the process risk, and possibly:
(2) a systematic component arising from parameter and model risk.

Process risk and parameter risk are clearly illustrated, in the previous Section, in terms of the variance of the residual lifetime T_x (see (5.8′)).

It is worth noting that, up to now, life insurance mathematics and survival modelling mostly focussed on process risk, with model and parameters taken as given. Non-life insurance mathematics has a quite different history, mainly thanks to Bayesian statistical models applied to inference about model parameters.

5.3 The longevity risk: actuarial evidence

In this Section, some examples of actuarial evidence of the longevity risk are provided. In particular, we focus on solvency requirements for life annuities portfolios (and funded pension plans). Solvency is investigated referring to immediate annuities, whence the so-called decumulation phase is only addressed. The material presented in this Section is drawn from Olivieri and Pitacco (2003), which the reader should refer to for the underlying algebra and further details and examples.

It should be stressed that the longevity risk also affects living benefits other than life annuities, e.g. long term care benefits, or lifetime sickness covers. See, for example, Olivieri and Pitacco (2002), Pitacco (2002).

5.3.1 Data, hypotheses and goals

We model life annuitants' mortality by assuming

$$\frac{q_x}{p_x} = G\,H^x \tag{5.10}$$

Note that the right-hand side of (5.10) is the third term in the Heligman-Pollard law, i.e. the term describing the old-age pattern of mortality (see (3.7)). In particular, G expresses the level of senescent mortality and H the rate of increase of senescent mortality itself. According to the notation used in Section 5.1 (see (5.2′) in particular), the parameter θ is represented by the couple (G, H). From (5.10) the survival function $S(x)$ can be obtained.

In order to represent mortality trends, we use projected survival functions. Seven survival functions $S^{[j]}(x)$, $j = 1, 2, \ldots, 7$, have been defined, to describe different mortality scenarios. The relevant parameters are given in Table 5.1. We point out that the scenarios differ one from the other in terms of the way they represent the phenomena of rectangularization and expansion (for both aspects, levels increase moving from scenario [1] to [7]).

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
G	0.000178	0.000042	0.000009	0.000002	0.0000004	0.0000001	0.00000001
H	1.07968	1.09803	1.11670	1.13450	1.15379	1.17215	1.19208

Table 5.1 – Parameters of the survival functions $S^{[j]}(x)$

As far as mortality risks are concerned, we adopt the following approaches:

(1) a "deterministic" approach, implemented by using a given projected survival function, chosen out of the set represented by Table 5.1; this approach only allows for the random fluctuation risk;

(2) a "stochastic" approach, implemented by using a set of projected survival functions, describing the uncertainty inherent in the projection; a "degree of belief" will be assigned to each function; this approach allows for both the random fluctuation risk and the longevity risk.

Let us focus on capital allocation. Following a traditional approach to solvency, meeting random liabilities is meant as the ability to set up the reserve for each policy in the portfolio and hence the portfolio reserve. Then, the insurance company is able to meet its liabilities at a given time t if and only if assets, Z_t, are greater than or equal to the portfolio reserve, $\mathcal{V}_t$, i.e., setting $M_t = Z_t - \mathcal{V}_t$, if and only if

$$M_t \geq 0 \tag{5.11}$$

In this framework, M_t is usually referred to as the solvency margin assigned to the portfolio.

To single out the effect of random mortality, we assume a deterministic financial structure, so that assets grow up according to a given interest rate structure.

Let 0 denote the time at which solvency is ascertained and assume $M_0 \geq 0$. Given a time horizon of T years, we say that the insurer has a solvency degree $1 - \varepsilon$ if and only if

$$\Pr\left\{\bigwedge_{t=1}^{T} M_t \geq 0\right\} = 1 - \varepsilon \tag{5.12}$$

Specific solvency requirements can be found by choosing proper values for the probability ε (the "ruin probability") and the time span T. Further, it is necessary to state whether capital flows affect the amount of assets Z_t, and to assume some hypotheses about portfolio future evolution. In what follows, we assume that no capital flow affects the portfolio, with the obvious exception of the capital allocation at time 0. Thus, the path of Z_t is only determined by benefits, interests and expenses. Moreover, a closed portfolio is addressed. In particular, the portfolio we are investigating consists of identical annuities, paid to persons of initial age $y = 65$, with annual amount $R = 100$. We assume that the future lifetimes of the annuitants have a common distribution and are independent of each other (conditional on any given survival function). The single premium (to be paid at entry) is calculated, for each policy, according to the survival function $S^{[4]}(x)$ and with a constant annual interest rate $i = 0.03$. Further, we assume that for each policy in force at time t, $t = 0, 1, \ldots$, a reserve must be set up, which is calculated according to such hypotheses. For further details (and for a different approach to solvency, as well), the reader should refer to Olivieri and Pitacco (2003).

Our aim is to determine the solvency margin required at time 0 such that, for a given choice of the quantities mentioned above, condition (5.12) is satisfied. We denote with M_0^* the required solvency margin at time 0.

5.3.2 A deterministic approach to mortality risks

In the deterministic approach to mortality, the probability distribution of the future lifetime of each insured is known, the only cause of uncertainty consisting in the time of death. The assessment of the required solvency margin is performed through simulation. In order to obtain results easier to interpret, we disregard profit; the actual life duration of the annuitants is thus simulated with the survival function $S^{[4]}(x)$. Further, we assume that the yield from investments is equal to the guaranteed rate of interest $i = 0.03$.

In Table 5.2 the solvency margin required at time 0, according to condition (5.12) with $\varepsilon = 0.025$, is plotted (in absolute and relative terms) against the initial portfolio size N_0. The time spans $T = \omega - y$ (where ω denotes the maximum age) and $T = 5$ years are alternatively chosen. Table 5.2 shows that the required margin decreases as N_0 increases. This is due to the fact that a deterministic approach to mortality only captures the risk of random fluctuations which, as it is well-known, is a pooling risk (i.e. its effect in respect of each policy decreases when larger numbers of similar policies are dealt with). Finally note that, since by definition the initial reserve $\mathcal{V}_0$ is equal to the single premium, the solvency margin at time 0 must be financed with shareholders' funds.

	$T=\omega-y$		$T=5$	
N_0	M_0^*	$\frac{M_0^*}{\mathcal{V}_0}$	M_0^*	$\frac{M_0^*}{\mathcal{V}_0}$
1000	32,959	2.180 %	17,242	1.140 %
2000	44,937	1.486 %	25,253	0.835 %
3000	54,398	1.199 %	27,771	0.612 %
4000	65,460	1.082 %	33,340	0.551 %
5000	73,242	0.969 %	36,621	0.484 %
6000	79,956	0.881 %	39,597	0.437 %
7000	87,585	0.828 %	44,861	0.424 %
8000	90,332	0.747 %	47,302	0.391 %
9000	100,555	0.739 %	49,744	0.366 %
10000	103,149	0.682 %	52,032	0.344 %

Table 5.2 – Required solvency margin (deterministic approach); ε=0.025

5.3.3 A stochastic approach to mortality risks

The assessment of the solvency margin is now obtained considering explicitly uncertainty in future mortality trends. To this aim, we consider the survival functions $S^{[j]}(x)$, $j = 1, 2, \dots, 7$, assuming that each of them is meant as a possible probability distribution of the future lifetime. A weight g_j, $j = 1, 2, \dots, 7$, is assigned to each survival function, representing the relevant "degree of belief".

The single premium for each policy and the individual reserve are still calculated with the survival function $S^{[4]}(x)$ and the interest rate $i = 0.03$. Two different sets of weights are assumed. The first set actually addresses just three scenarios and is as follows:

$$g_2 = g_6 = 0.20;\; g_4 = 0.60;\; g_1 = g_3 = g_5 = g_7 = 0 \tag{5.13a}$$

whereas the second set is as follows:

$$g_1 = g_7 = 0.01;\; g_2 = g_6 = 0.04;\; g_3 = g_5 = 0.20;\; g_4 = 0.50 \tag{5.13b}$$

Note that both choices reflect the fact that $S^{[4]}(x)$, which is used for pricing and reserving, is assumed to provide the most reliable scenario description.

Obviously, the investigation is carried out through simulation. We now deal with two causes of uncertainty: the actual distribution of future lifetimes and the age at death of each annuitant. Firstly, the survival function must be chosen (through simulation) and then, assuming that under a given lifetime distribution the annuitants constitute independent risks, the actual duration of life of each person is simulated.

Results in Table 5.3 are based on the set of weights (5.13a). For any portfolio size, the comparison between the deterministic and the stochastic approach shows a heavy increase of the required solvency margin in the latter. This is due to the fact that a stochastic framework allows us to analyze not only the risk of random fluctuations in the number of survivors around the relevant expected value, but also that of systematic deviations (i.e. the longevity risk), which is a non-pooling risk. The (slightly) decreasing behaviour of the relative required solvency margin with respect to N_0 is due to the pooling effect concerning random fluctuations; however, its magnitude is rather stable and its value seems to tend to a large positive amount; hence, the non-pooling effect of longevity risk is witnessed.

Assuming $T = 5$ as the time horizon leads to a solvency margin significantly lower than assuming $T = \omega - y$. This shows that the longevity risk reveals itself in the long run. Actually, the choice $T = 5$ implies a strong postponement of the solvency margin building up and the need to monitor carefully the portfolio in order to adjust the solvency margin in case it is insufficient.

From results in Table 5.3, pertaining to $T = \omega - y$, it could be argued that when just three scenarios are actually dealt with (see the weights in (5.13a)), the dramatic increase in the solvency margin when considering also the longevity risk is due to the difference between the reserve calculated with the "worst" assumed scenario and the reserve based on the central one. Denoting by $V_t^{[\cdot]}$ the individual reserve calculated with the survival function $S^{[\cdot]}(x)$, we find

$$\frac{V_0^{[6]}}{V_0^{[4]}} - 1 \;=\; 13.383\%$$

So, in the example the magnitude of the difference between such reserves is actually similar to the magnitude of the required solvency margin.

However this example does not lead to any general conclusion, given that the individual reserve is an expected value of liabilities, whilst the solvency reserve is related to the right tail of the distribution of assets. In order to have a better understanding, let us consider a wider set of scenarios (see (5.13b)). Table 5.4 shows the relevant results, corresponding to some values of the ruin probability.

As far as a worst case analysis is concerned, it is worth noting that, for the individual reserve, we now find

$$\frac{V_0^{[7]}}{V_0^{[4]}} - 1 \;=\; 20.631\%$$

scenario [7] representing the worst case. Hence a worst case reserving clearly leads to a huge reserve. More generally, setting aside a solvency margin simply based on the comparison among reserves calculated with different survival functions (as some practice suggests) on the one hand would disregard the risk of random fluctuations (which obviously can be considered separately) and on the other would disregard a valuation of the probability of ruin, possibly leading to unsound capital allocation.

Results and remarks about solvency requirements and reserving can be summarized as follows. Assume that the portfolio reserve $\mathcal{V}_0$ is calculated using a "best estimate" basis as far as mortality is concerned, e.g. $S^{[4]}(x)$. Let $\mathcal{V}_0^{[W]}$ and $\mathcal{V}_0^{[B]}$ denote respectively the portfolio reserve calculated according to the "worst

N_0	T=$\omega-y$ M_0^*	$\frac{M_0^*}{\mathcal{V}_0}$	T=5 M_0^*	$\frac{M_0^*}{\mathcal{V}_0}$
1000	218,172	14.431 %	43,869	2.902 %
2000	425,758	14.081 %	83,771	2.770 %
3000	632,324	13.941 %	123,672	2.727 %
4000	840,797	13.903 %	162,592	2.689 %
5000	1,046,066	13.838 %	202,827	2.683 %
6000	1,253,963	13.824 %	241,871	2.666 %
7000	1,457,367	13.771 %	279,694	2.643 %
8000	1,660,461	13.729 %	318,527	2.634 %
9000	1,867,065	13.722 %	357,704	2.629 %
10000	2,072,830	13.710 %	395,813	2.618 %

Table 5.3 – Required solvency margin (stochastic approach); ε=0.025

ε	T=$\omega-y$ M_0^*	$\frac{M_0^*}{\mathcal{V}_0}$	T=5 M_0^*	$\frac{M_0^*}{\mathcal{V}_0}$
0.025	206,852	13.682 %	39,592	2.619 %
0.05	156,250	10.335 %	33,531	2.218 %
0.1	108,272	7.161 %	26,871	1.777 %

Table 5.4 – Required solvency margin (stochastic approach); N_0=1000

case" basis, e.g. $S^{[7]}(x)$, and a "bad case" basis, e.g. $S^{[5]}(x)$ or $S^{[6]}(x)$. Clearly, we have:

$$\mathcal{V}_0 < \mathcal{V}_0^{[B]} < \mathcal{V}_0^{[W]} \tag{5.14}$$

Further, let us define the following ratios (where ${M_0^*}^{det}$ and ${M_0^*}^{prob}$ refer to the deterministic and the stochastic approach respectively):

$$Q^{det}(N_0, \varepsilon) = \frac{{M_0^*}^{det}}{\mathcal{V}_0} \tag{5.15a}$$

$$Q^{prob}(N_0, \varepsilon) = \frac{{M_0^*}^{prob}}{\mathcal{V}_0} \tag{5.15b}$$

$$R^{[\cdot]} = \frac{\mathcal{V}_0^{[\cdot]}}{\mathcal{V}_0} - 1 \tag{5.15c}$$

Ratios (5.15a) and (5.15b) are functions of the (initial) portfolio size N_0 and the probability ε, whilst the ratio (5.15c), for any choice of the reserving basis, is independent of both N_0 and ε as it does not allow for any mortality risk.

From (5.14) we have:

$$0 < R^{[B]} < R^{[W]} \tag{5.16}$$

whereas numerical results show that, obviously:

$$Q^{det}(N_0, \varepsilon) < Q^{prob}(N_0, \varepsilon) \tag{5.17}$$

Of course, no general conclusion can be reached as regards the relationship between ratios R's and Q's. Notwithstanding, a likely situation is sketched by Figure 5.1.

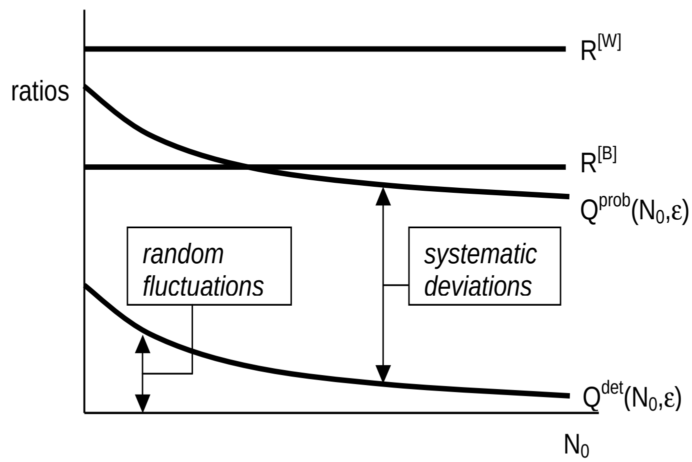

Figure 5.1 – Ratios R's and Q's

Numerical results only provide an illustration, since they obviously depend on the specific assumptions concerning randomness in mortality. Nevertheless, the results presented in this Section underline the dramatic importance of a sound evaluation of the solvency requirements for life annuities (and, more generally, for insurance products providing lifetime living benefits, e.g. long term care covers, post-retirement sickness benefits, etc.).

6. CONCLUSIONS

Evolution of survival modelling in life insurance mathematics has been analyzed, and several features of this evolution have been singled out. How do the various features enter the actuarial models? The following aspects should be considered.

(a) In some cases "new" variables, i.e. variables other than the attained age, enter the model. For example:

– the attained age $x + t$ must be split into the two terms x (age at policy issue) and t (policy duration) when issue-selection is allowed for (see Section 3.4.1);

– allowance for risk factors implies the presence of a number of parameters in pricing formulae (see Section 3.4.2);

– the calendar year y enters the model when mortality trends are accounted for (see Sections 4.2 and 4.3).

(b) An extended structure is used when also benefits depending on random variables different from the insured's lifetime are concerned. So, multiple decrements models and multistate models are used, rather than the traditional two-state model, for example when disability benefits are included in the insurance cover (see Section 3.3).

(c) Allowing for risk requires new outputs from the actuarial model (and more detailed inputs as well). For instance, variances, percentiles, etc. of some probability distribution provide information about the riskiness of an insurance policy or a portfolio (see Sections 3.1 and 5.3).

(d) Premium principles other than the (traditional) equivalence principle should be used when pricing guarantees (and possibly options) embedded in insurance products. This is the case of annuity products in which the annuity amount is guaranteed whatever the mortality trend may be. Pricing this guarantee means that the policyholder is charged with part of the cost of solvency (see Section 5.3).

It is worth noting that premium principles different from the equivalence principle are rather popular in non-life insurance mathematics. This fact suggests some comparisons between the evolution of actuarial modelling and practice in life and, respectively, non-life insurance.

As far as actuarial practice is concerned, the role of the actuary belongs to the tradition of life insurance. In fact, it can be traced back to the 1750's as witnessed, for instance, by James Dodson's work (see Section 2.2). Non-life insurance history probably began in northern Italy about the end of the twelfth century, with marine insurance. Despite this long history, as Haberman (1996) notes, the list of actuarial contributions to this area of practice is much briefer. As in the life insurance field, the earliest contributions to the actuarial practice of non-life insurance date back to the half of 17th century. Corbyn Morris in 1747 showed how an increasing portfolio size reduces the risk of an insurer, and this work probably constitutes one of the first attempts to use mathematics in dealing with practical non-life insurance problems (see Haberman, 1996). Nevertheless, actuarial mathematics gains an important role in non-life insurance much later. The credibility formula, proposed by A. Whitney in 1918, probably represents the first important mathematical contribution to non-life actuarial practice.

As regards actuarial theory, it should be recognized that a number of important aspects have been analyzed in the framework of non-life insurance mathematics and later addressed in the life insurance context. Two meaningful examples follow.

(1) The assumption of heterogeneity is quite natural in non-life insurance mathematics. In particular, unobservable heterogeneity factors (see Section 3.4.3) are allowed for in actuarial models describing the random number of claims in a given period, through a random parameter $\tilde{\theta}$ included in the model itself. A probability distribution is then assigned to $\tilde{\theta}$, for example via a density function $g(\theta)$. Probably, the most popular implementation of this structure is given by the Poisson process with random parameter, described by a gamma distribution.

This implementation leads to the well-known negative binomial distribution for the random number of claims. Insured risks, assumed to be independent under any given hypothesis $\tilde{\theta} = \theta$, are unconditionally correlated. An analogous correlation concerns lifetimes of policyholders in the life insurance framework, when unobservable heterogeneity is accounted for.

(2) Risks other than process risk entered the life insurance actuarial theory very recently, in particular because of the presence of longevity risk (see Section 5.2). In a non-life insurance context, the awareness of further types of risk emerges, for instance, from the adoption of pricing procedures in group life insurance (see for example Norberg, 1989) aiming to "learn" from experience. In this sense, credibility models can be seen as tools to keep the parameter risk at reasonable levels. In a different perspective, the adoption of credibility models proves the awareness of heterogeneity within the insured group.

REFERENCES

Amsler M H (1968), Les chaînes de Markov des assurances vie, invalidité et maladie, *Transactions of the 18th International Congress of Actuaries*, München, vol. 5: 731–746

Beard R E (1952), Some further experiments in the use of the incomplete gamma function for the calculation of actuarial functions, *Journal of the Institute of Actuaries*, **78**: 341–353

Beard R E (1959), Note on some mathematical mortality models, in: C E W Wolstenholme and M O Connor (eds), *The lifespan of animals*, CIBA Foundation Colloquia on Ageing, vol. 5, Boston: 302–311

Beard R E (1971), Sone aspects of theories of mortality, cause of death analysis, forecasting and stochastic processes, in: W Brass (ed), *Biological Aspects of Demography*, Taylor and Francis, London: 57–68

Benjamin B and J H Pollard (1993), *The analysis of mortality and other actuarial statistics*, The Institute of Actuaries, Oxford

Benjamin B and A S Soliman (1993), *Mortality on the move*, Actuarial Education Service, Oxford

Blaschke E (1923), Sulle tavole di mortalità variabili col tempo, *Giornale di Matematica Finanziaria*, **5**: 1–31

Bourgeois-Pichat J (1952), Essai sur la mortalité "biologique" de l'homme, *Population*, **7** (3): 381-394

Brass W (1974), Mortality models and their uses in demography, *Transactions of the Faculty of Actuaries*, **33**: 123–132

Brouhns N and M Denuit (2002), Risque de longévité et rentes viagères. II. Tables de mortalité prospectives pour la population belge, *Belgian Actuarial Bulletin*, **2**: 49–63

Buettner T (2002), Approaches and experiences in projecting mortality patterns for the oldest-old, *North American Actuarial Journal*, **6** (3): 14–25

Butt Z and S Haberman (2002), *Application of frailty-based mortality models to insurance data*, Actuarial Research Paper No. 142, Dept. of Actuarial Science and Statistics, City University, London

Cairns A J G (2000), A discussion of parameter and model uncertainty in insurance, *Insurance: Mathematics & Economics*, **27** (3): 313–330

CMIR10 (Continuous Mortality Investigation Reports) (1990), Institute of Actuaries and Faculty of Actuaries

CMIR17 (Continuous Mortality Investigation Reports) (1999), Institute of Actuaries and Faculty of Actuaries

Cramér H and H Wold (1935), Mortality variations in Sweden: a study in graduation and forecasting, *Skandinavisk Aktuarietidskrift*, **18**: 161–241

Cummins J D, B D Smith, R N Vance and J L VanDerhei (1983), *Risk classification in life insurance*, Kluwer-Nijhoff Publishing

Daboni L (1964), Modelli di rendite aleatorie, *Giornale dell'Istituto Italiano degli Attuari*, **27**: 273–296

Davidson A R and A R Reid (1927), On the calculation of rates of mortality, *Transactions of the Faculty of Actuaries*, **11** (105): 183–232

Daw R H (1979), Smallpox and the double decrement table. A piece of actuarial pre-history, *Journal of the Institute of Actuaries*, **106**: 299–318

de Finetti B (1950), Matematica attuariale, *Quaderni dell'Istituto per gli Studi Assicurativi*, Trieste, **5**: 53–103

de Finetti B (1957), *Lezioni di matematica attuariale*, Edizioni Ricerche, Roma

Forfar D O and D M Smith (1988), The changing shape of English Life Tables, *Transactions of the Faculty of Actuaries*, **40**: 98–134

Franckx E (1963), Essai d'une théorie opérationelle des risques markoviens, *Quaderni dell'Ist. di Matematica Finanziaria dell'Università di Trieste*, No. 11

Graf G (1906), *Il funzionamento matematico delle Assicurazioni Generali in Trieste*, Assicurazioni Generali, Trieste

Haberman S (1996), *Landmarks in the history of actuarial science (up to 1919)*, Actuarial Research Paper No. 84, Dept. of Actuarial Science and Statistics, City University, London

Haberman S and E Pitacco (1999), *Actuarial models for disability insurance*, Chapman & Hall / CRC Press, London

Haberman S and T A Sibbett (eds) (1995), *History of actuarial science*, Pickering & Chatto, London

Hald A (1987), On the early history of life insurance mathematics, *Scandinavian Actuarial Journal*: 4–18

Heligman L and J H Pollard (1980), The age pattern of mortality, *Journal of the Institute of Actuaries*, **107**: 49–80

Hoem J M (1969), Markov chain models in life insurance, *Blätter der deutschen Gesellschaft für Versicherungsmathematik*, **9**: 91–107

Hoem J M (1972), Inhomogeneous semi-Markov processes, select actuarial tables and duration-dependence in demography, in: Greville T N E (ed.), *Population Dynamics*, Academic Press, London: 251–296

Hoem J M (1988), The versatility of the Markov chain as a tool in the mathematics of life insurance, *Transactions of the 23-rd International Congress of Actuaries*, Helsinki, vol. R: 171–202

Horiuchi S and J R Wilmoth (1998), Deceleration in the age pattern of mortality at older ages, *Demography*, **35** (4): 391–412

Keyfitz N (1982), Choice of functions for mortality analysis: Effective forecasting depends on a minimum parameter representation, *Theoretical Population Biology*, **21**: 329–352

Keyfitz N (1985), *Applied mathematical demography*, Springer-Verlag, New York

Kopf E W (1926), The early history of life annuity, *Proceedings of the Casualty Actuarial Society*, **13** (27): 225–266

Lee R D (2000), The Lee-Carter method for forecasting mortality, with various extensions and applications, *North American Actuarial Journal*, **4** (1): 80–93

Lee R D and L R Carter (1992), Modelling and forecasting U.S. mortality, *Journal of the American Statistical Association*, **87** (14): 659–675

Lemaire J and A S MacDonald (2003), *Genetics, family history, and insurance underwriting: an expensive combination?*, Invited Lecture at the 34th ASTIN Colloquium, Berlin (www.astin2003.de)

Levinson L H (1959), A theory of mortality classes, *Transactions of the Society of Actuaries*, **11**: 46–87

MacDonald A S (1999), Modelling the impact of genetics on insurance, *North American Actuarial Journal*, **3** (1): 83-105

MacDonald A S, A J G Cairns, P L Gwilt and K A Miller (1998), An international comparison of recent trends in population mortality, *British Actuarial Journal*, **4**: 3–141

Møller C M (1990), Select mortality and other durational effects modelled by partially observed Markov chains, *Scandinavian Actuarial Journal*: 177–199

Norberg R (1988), Select mortality: possible explanations, *Transactions of the 23-rd International Congress of Actuaries*, Helsinki, vol. 3: 215–224

Norberg R (1989), Experience rating in group life insurance, *Scandinavian Actuarial Journal*: 194–224

Nordenmark N V E (1906), Über die Bedeutung der Verlängerung der Lebensdauer für die Berechnung der Leibrenten, *Transactions of the 5th International Congress of Actuaries*, Berlin, vol. 1: 421–430

Olivieri A (2001), Uncertainty in mortality projections: an actuarial perspective, *Insurance: Mathematics & Economics*, **29** (2): 231–245

Olivieri A and E Pitacco (2002), Premium systems for post-retirement sickness covers, *Belgian Actuarial Bulletin*, **2**: 15–25

Olivieri A and E Pitacco (2003), Solvency requirements for pension annuities, to appear on: *Journal of Pension Economics & Finance*

Petrioli L and M Berti (1979), *Modelli di mortalità*, F. Angeli Editore, Milano

Pitacco E (2002), *Longevity risks in living benefits*, Working Paper CeRP No. 23/02, Moncalieri (Torino) (to appear on: E Fornero, E Luciano (eds.), *Developing an annuity market in Europe*, Edward Elgar, Cheltenham)

Pollard A H (1949), Methods of forecasting mortality using Australian data, *Journal of the Institute of Actuaries*, **75**: 151–182

Pollard A H (1970), Random mortality fluctuations and the binomial hypothesis, *Journal of the Institute of Actuaries*, **96**: 251–264

Pollard J H (1987), Projection of age-specific mortality rates, *Population Bulletin of the UN*, **21-22**: 55–69

Poulin C (1980), Essai de mise au point d'un modèle représentatif de l'évolution de la mortalité humaine, *Transactions of the 21st International Congress of Actuaries*, Zürich-Lausanne, vol. 2: 205–211

Redington F M (1969), An exploration into patterns of mortality, *Journal of the Institute of Actuaries*, **95**: 243–298

Renshaw A E and S Haberman (2000), *Modelling for mortality reduction factors*, Actuarial Research Paper No. 127, Dept. of Actuarial Science and Statistics, City University, London

Renshaw A E and S Haberman (2003a), On the forecasting of mortality reduction factors, *Insurance: Mathematics and Economics*, **32** (3): 379–401

Renshaw A E and S Haberman (2003b), Lee - Carter mortality forecasting, a parallel GLM approach, England & Wales mortality projections, *Applied Statistics*, **52**: 119–137

Rüttermann M (1999), Mortality trends worldwide, *Risk Insights*, General & Cologne RE, **3** (4): 18–20

Seal H L (1977), Studies in the history of probability and statistics. XXXV. Multiple decrements or competing risks, *Biometrika*, **64**: 429–439

Sithole T Z, S Haberman and R J Verrall (2000), An investigation into parametric models for mortality projections, with applications to immediate annuitants and life office pensioners' data, *Insurance: Mathematics and Economics*, **27** (3): 285–312

Smith D and N Keyfitz (1977), *Mathematical demography. Selected papers*, Springer-Verlag, Berlin

Smolensky P (1934), Sulle tavole compatte di mortalità, *Transactions of the 10th International Congress of Actuaries*, Rome, vol. 2: 235–251

Sverdrup E (1952), Basic concepts in life assurance mathematics, *Skandinavisk Aktuarietidskrift*, **3-4**: 115–131

Tabeau E, A van den Berg Jeths and C Heathcote (eds) (2001), *Forecasting mortality in developed countries*, Kluwer Academic Publishers

Vaupel J W, K G Manton and E Stallard (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, **16** (3): 439–454

Wetterstrand W H (1981), Parametric models for life insurance mortality data: Gompertz's law over time, *Transactions of the Society of Actuaries*, **33**: 159–175

Wilmoth J R and S Horiuchi (1999), Rectangularization revisited: variability of age at death within human populations, *Demography*, **36** (4): 475–495