

Appendix I

Uncertainty analysis in longitudinal population models

Introduction¹

Multi-state longitudinal life tables are used to measure population health as well as to study the effect of specific interventions implemented for the sake of improving population health (Lauer et al. 2003). How does uncertainty in parameter values used as inputs affect the projections of multi-state longitudinal life tables? It seems that the uncertainty in the results of such models will depend on the individual contributions of multiple uncertain parameters, as well as on the dynamic characteristics of the model itself.

Despite their obvious interest, these and related questions seem to have been ignored in the relevant literature on demography, ecology (e.g. matrix population models; Caswell, 2001) and epidemiology. In fact, to identify methods relevant to the analysis of these questions, one has to turn to the mathematical, physical and statistical literature on stochastic processes and Markov models.

We shall discuss uncertainty in the rate parameters of the population model only. This restriction is partly for the sake of convenience and partly a reflection of the fact that rates of mortality and disease are greater determinants of population health than other parameters (e.g. the sizes, or relative sizes, of initial populations). Moreover, we shall not develop detailed results for the distributions of individual life table quantities but shall rather discuss the methods and principles that allow for a general approach to such calculations.

1 Harald Wirth, Klaus Röhrich, Jan Barendregt and Dik Habbema are gratefully acknowledged for their helpful discussions on this appendix.

Formal statement of the problem

For the purposes of population health measurement, we consider a differential equation system of the following form:

$$\text{Eq. (1)} \quad \mathbf{p}'(t) = \mathbf{C} \cdot \mathbf{p}(0),$$

where $\mathbf{p}' \equiv d\mathbf{p}/dt$ denotes the derivative of $\mathbf{p}$ with respect to time, $\mathbf{C}$ is a matrix of non-time-varying rates and $\mathbf{p}(0)$ is the vector of initial populations. For the case of the population model, it is natural to consider $\mathbf{p}(0)$ as the population of interest partitioned in the set of defined age groups.

It is well known that the general solution to the system of differential equations can be written as (see, for example, Strang, 1988):

$$\text{Eq. (2)} \quad \mathbf{p}(t) = e^{\mathbf{C} \cdot t} \cdot \mathbf{p}(0).$$

This result means that initial populations evolve exponentially with time; Eq. (2) is merely a concise way of representing the population model.

In terms of Eq. (2), we are interested in the uncertainty of $\mathbf{p}(t)$ when there is uncertainty in the elements of $\mathbf{C}$. We are also interested in the uncertainty of various life-table functions of the populations $\mathbf{p}(t)$, which we can write as $f(\mathbf{p}(t))$. Important examples of such functions are the average population size in an interval (population residence time) and the sums of populations or of average populations over a given horizon (Lauer et al., 2003).

Scope of the analysis

Uncertainty in $\mathbf{p}(t)$ can in general be attributed to the following sources:

- model uncertainty,
- stochastic uncertainty (i.e. random variation) in the modelled process, and
- uncertainty in model parameters.

We briefly describe each of these sources below.

Model uncertainty. Model uncertainty refers to uncertainty in the populations $\mathbf{p}(t)$ when the specific mathematical method used to model the population (e.g. Eq. (2)), or its related assumptions, are incorrect or represent a simplification or idealization of the true underlying population process. For example, exponential evolution in continuous time may be used to approximate an underlying process that is in fact discontinuous. Alternatively, rates may not be constant but time varying. Typically, however, model uncertainty arises when the matrix $\mathbf{C}$ is missing elements that are important for the question under study. This can arise due to the fact only a small number of known transitions are typically represented in $\mathbf{C}$, as well as to the fact that there are many important transitions that are unknown (e.g. exposure to unknown risk factors).

Stochastic uncertainty. Stochastic uncertainty arises when the population process is inherently stochastic. It is useful to distinguish the uncertainty arising from modelling a discrete process by means of a continuous one, or from modelling a random process by means of a deterministic one (both of which would be causes of model uncertainty in terms of the definition above), from the intrinsic variability of an inherently random process.

Parameter uncertainty. Parameter uncertainty refers to the fact that parameters in the population model may be unknown or are measured imperfectly. Measurement uncertainty can either be statistical or systematic. Statistical measurement uncertainty refers to random variation in either the measurement tools or in the measurers (which implies that measurement is a random process, i.e. subject to inherent random variation). Since statistical parameter uncertainty can either be model error (for example, if the measurement tool is deterministic but the process measured is random) or

simple stochastic uncertainty, it is therefore best considered under one or the other of those respective headings. On the other hand, systematic parameter uncertainty can arise due to the presence in the measurement tool of non-random dependence on important parameters whose values are unmeasured or unknown or to systematic bias in the measurement method (cf. Hubble telescope). Notice therefore that systematic parameter uncertainty can therefore be considered as a form of model uncertainty.

In terms of these definitions, the analysis here applies to uncertainty in model outcomes due to stochastic variability in the elements of the matrix of rates $\mathbf{C}$. However, the approach is also well suited to the characterization of statistical parameter uncertainty arising due to random errors in the measurement tool or in the measurers). Finally, inherently random processes modelled by deterministic processes will display model error that is assimilable to the sort of uncertainty characterized here. We do not address other forms of model uncertainty or systematic parameter uncertainty.

In the notation of (Lauer et al., 2003), the matrix of rates $\mathbf{C}$ can be written as follows:

$$\begin{pmatrix} h_{S \rightarrow X} + h_{S \rightarrow C} + m & -h_{X \rightarrow S} & -h_{C \rightarrow S} & 0 \\ -h_{S \rightarrow X} & h_{X \rightarrow XC} + h_{X \rightarrow S} + m + f_X & 0 & -h_{XC \rightarrow X} \\ -h_{S \rightarrow C} & 0 & h_{C \rightarrow XC} + h_{C \rightarrow S} + m + f_C & -h_{XC \rightarrow C} \\ 0 & -h_{X \rightarrow XC} & -h_{C \rightarrow XC} & h_{XC \rightarrow C} + h_{XC \rightarrow X} + m + f_X + f_C \end{pmatrix}$$

Note that, other than the stochastic component, all matrix entries are non-time varying. Transitions between health states are denoted by h , and distinguished by subscripts, and background mortality and case fatality rates are labelled m and f respectively, the latter also distinguished by subscripts. In terms of that notation, $\mathbf{p}(0) = (S(0) \ X(0) \ C(0) \ XC(0))^T$, where the population of each possible health state is defined at $t = 0$.

Analysis of specific cases

It is convenient to consider initially the case of a single population exposed to a single transition, in other words, the case of a standard cohort life table, where $C = \mu$ and $p(0) = p(0)$.

Cohort life table under uncertainty

We refer to a continuous-time life table comprising a single state exposed to constant mortality rate μ_y , where the subscript y indexes the rate prevailing during period y (notionally, one year). With subscripts now denoting time, within any period the population at time t is given by:

$$\text{Eq. (3)} \quad p_t = p_{t-1} e^{-\mu_y \cdot t}.$$

Suppose $p_{t-1} = p_0$ and $p_0 = 1$ (the latter condition can always be enforced without loss of generality by dividing both p_t and p_0 by p_0). Suppose, moreover, that $t = 1$ year and that μ_y has units “year⁻¹”. Then we can write:

$$\text{Eq. (4)} \quad p_1 = e^{-\mu_1}$$

for the population surviving at the end of year 1.

Now suppose that, starting from the beginning of year 2, the population surviving at the end of year 1 is exposed to μ_2 . Substituting the value obtained for p_1 , it follows that the population surviving at the end of year 2 can be written as:

$$\text{Eq. (5)} \quad p_2 = e^{-\mu_1} \cdot e^{-\mu_2} = e^{-(\mu_1 + \mu_2)}.$$

Similarly, if the population surviving at the end of year 2 is exposed to μ_3 from the beginning of year 3, the surviving population at the end of year 3 is given by:

$$\text{Eq. (6)} \quad p_3 = e^{-(\mu_1 + \mu_2)} \cdot e^{-\mu_3} = e^{-(\mu_1 + \mu_2 + \mu_3)}.$$

Under these assumptions, it is clear that in year t ,

$$\text{Eq. (7)} \quad p_t = e^{-\sum_{y=1}^t \mu_y}.$$

Notice, however, that these assumptions are merely those of the cohort version of the standard life table in continuous time where age-specific mortality is assumed constant within an age group but differs by year of age.

Independent rates

Now suppose that μ_y no longer represents a constant mortality rate but is instead the mean value of a normally distributed random rate h_y , with variance σ_y^2 , and that the h_y are independent. We refer to this concept of random variability as the “null model”. The null model provides a useful reference case for comparison with more complicated types of variability.

It follows from the definition of the log-normal random variable that P_1 is not a constant with value $e^{-\mu_1}$ but rather a log-normally distributed random variable with mean $e^{-\mu_1 + \frac{1}{2}\sigma_1^2}$ and variance $e^{-2\mu_1 + 2\sigma_1^2} - e^{-2\mu_1 + \sigma_1^2}$, the latter equal to $e^{-2\mu_1 + \sigma_1^2}(-1 + e^{\sigma_1^2})$ after factoring. Notice that, when compared with the case of constant rates, mean mortality is reduced by the quantity $\frac{1}{2}\sigma_y^2$. Regardless of the precise distributional form, when rates are random the mean population at year's end will show correction factors whose values are subtracted from mean mortality. Thus, the mean number of survivors when rates are random is greater than the number of survivors when rates are constant.

The sum of t independent normal random variables with means μ_y and variances σ_y^2 is itself a normally distributed random variable with mean $\sum_t \mu_y$ and $\sum_t \sigma_y^2$ (Gardiner, 2004). It follows that p_t is a log-normally distributed random variable with mean $e^{\sum_t \mu_y + \frac{1}{2}\sum_t \sigma_y^2}$ and variance $e^{-2(\sum_t \mu_y + \frac{1}{2}\sum_t \sigma_y^2) + \sum_t \sigma_y^2} + e^{-2(\sum_t \mu_y + \frac{1}{2}\sum_t \sigma_y^2) + 2\sum_t \sigma_y^2}$, the latter expression equal to $e^{-2\sum_t \mu_y}(-1 + e^{\sum_t \sigma_y^2})$ after cancellations and factoring.

Since the distributional family of populations in the cohort null model is invariant by year, it is natural to think of the quantities $\sum_t \mu_y$ as “pseudorates” (i.e. “cumulative rates”) with variance $\sum_t \sigma_y^2$. Since pseudorates are in this respect indistinguishable from actual (annual) rates, uncertainty analysis of the cohort null model can be greatly simplified. For example, if all annual rates have mean μ and variance σ^2 , the mean population after t years is simply $e^{\mu + \frac{1}{2}t\sigma^2}$, and the corresponding expression for population variance is $e^{-2t\mu}(-1 + e^{t\sigma^2})$.

Population correlations when rates are independent

Although rates are independent in the cohort null model, age-specific populations follow a specific correlation law. Write $h_y = \mu_y + dh_y$, where dh_y is a normally distributed random variable with a mean of zero and variance σ_y^2 . We know from the population model that:

$$\text{Eq. (8)} \quad p_t = p_{t-1} e^{-h_y}.$$

Rewriting h_y in terms of deviations around its mean, we can moreover write:

$$\text{Eq. (9)} \quad e^{-h_y} = e^{-\mu_y - dh_y}.$$

By both adding and subtracting the quantity $\frac{1}{2}\sigma_y^2$ in the exponent and factoring we then write:

$$\text{Eq. (10)} \quad e^{-h_y} = e^{-\mu_y + \frac{1}{2}\sigma_y^2} e^{-dh_y - \frac{1}{2}\sigma_y^2}.$$

It follows that the population model can be written as:

$$\text{Eq. (11)} \quad p_t = p_{t-1} e^{-\mu_y + \frac{1}{2}\sigma_y^2} e^{-dh_y - \frac{1}{2}\sigma_y^2}.$$

But Eq. (11) says that p_t is equal to the product of p_{t-1} , a positive constant $e^{-\mu_y + \frac{1}{2}\sigma_y^2}$ (whose value is less than 1), and a random variable $e^{-dh_y - \frac{1}{2}\sigma_y^2}$ with a mean of 1 (since $-dh - \frac{1}{2}\sigma_y^2$ is normally distributed with mean $-\frac{1}{2}\sigma_y^2$ and variance σ_y^2 , its exponential is a log-normal random variable with mean $e^{-\frac{1}{2}\sigma_y^2 + \frac{1}{2}\sigma_y^2}$, which is equal to e^0).

A closely related model from the time-series literature is the first-order autoregressive (AR-1) process (Goldberger, 1991):

$$\text{Eq. (12)} \quad p_t = \alpha p_{t-1} + \varepsilon_t,$$

where ε_t is a random variable with zero mean and variance σ^2 and α is a constant whose value is less than 1. In a similar notation, the cohort null model could be written:

$$\text{Eq. (13)} \quad p_t = \alpha_y p_{t-1} \xi_t,$$

where α_y is equal to $e^{-\mu_y + \frac{1}{2}\sigma_y^2}$ and ξ_t is equal to $e^{-dh_y - \frac{1}{2}\sigma_y^2}$.

With identically distributed rates, by induction on t it can be seen that the covariance of two populations t years different in age is given by $\alpha' \sigma^2$; this is clearly the same correlation law as for the additive AR-1 process (Goldberger, 1991), justifying the application of the term “first-order autoregressive” to the null cohort model. Moreover, when rates differ by year, it is not hard to see that the covariance of two populations separated by t years is given, by induction on t , by the variance of the first population σ_1^2 times the product of the t coefficients α_y , in other words, $\sigma_1^2 \cdot \prod_t \alpha_y$. The error process of the cohort null model is thus a multiplicative analog of the standard AR-1 process, with the additional possibility that α_y is a period-specific constant.

Such a model is also closely related to geometric Brownian motion and to the geometric random walk (Franke et al., 2004). Taking the logarithm of both sides of Eq. (11), we get:

$$\text{Eq. (14)} \quad \log p_t = \log p_{t-1} - \mu_y - dh_y.$$

But, since dh_y has a mean of zero, we can also write:

$$\text{Eq. (15)} \quad \log p_t = -\mu_y + \log p_{t-1} + dh_y.$$

Thus, the logarithm of population in the cohort null model follows a random walk with a period-specific drift term $-\mu_y$. This differs from the standard geometric random walk in that the drift term can be period specific; it differs from geometric Brownian motion in that dh_y is not continuously random.

In sum, the correlation law for populations in the cohort null model is well defined. The null model correlation law can also be related to well known stochastic-process and time-series models.

Distributional characteristics of residence time

So far we have focused on the uncertainty analysis of population quantities. The population residence time (i.e. aggregate population life years, p_{LY}) in an interval is another quantity of interest.

Population residence time is equal to the time integral of the population within the interval. For example, in the cohort null model for the first year, we have (Lauer et al., 2003):

$$\text{Eq. (16)} \quad p_{LY(1)} = \int_0^1 p_0 e^{-h_1 t} dt = \frac{1 - e^{-h_1}}{h_1} \cdot p_0.$$

For h_1 random, the right-hand side of Eq. (16) defines a distribution for $p_{LY(1)}$. Although p_0 is non-stochastic by assumption, in general p_{t-1} is a random variable. Nevertheless, we focus first on the distributional characteristics of the expression $(1 - e^{-h})/h$. In a continuous-time life table, this expression gives the general formula for the time average of survivors of a population exposed to mortality rate h during a one-year interval. Since the time average of survivors is nothing other than the population residence time (life years) accumulated during the interval, the distribution defined by Eq. (16) is of general interest for population health measurement.

Since the distributional characteristics of residence time cannot in general be given in terms of closed-form analytical expressions (exceptions are when h is gamma or uniform distributed), we seek approximate expressions for its mean and variance. For any standard distributions for h , approximations can be found by means of the delta method (Goldberger, 1991). For the mean, the delta method to linear order implies taking the function of the mean $f(\mu)$ as an approximation to the mean of the function $E[f(h)]$, in other words we take the quantity $(1 - e^{-\mu})/\mu$ as an approximation to the mean of $(1 - e^{-h})/h$.

To obtain the delta-method approximate variance, the Taylor-series expansion of the function around the mean of h is truncated at linear order, and the theorem for variance of a linear function is used. This yields the expression $(e^{-2\mu}(1 - e^{\mu} + \mu)^2)\mu^{-4}$ as the linear-order delta-method approximation to the variance of $(1 - e^{-h})/h$. Thus, linear-order approximations for both the mean and the variance of residence time are expressible in terms of the mean only of the rate h .

If it is relevant to consider the value of p_{t-1} as a random variable, residence time in the interval is therefore the product of two random variables. Calculating the expectation and variance of the product of random variables requires the covariance of the random variables. By using the delta-method result for the mean (i.e. $(1 - e^{-\mu})/\mu$), the required covariance can be given as approximately $-(e^{-\mu_y + \frac{1}{2}\sigma_y^2})\mu^{-2}\sigma^2$.

Correlated rates

We maintain the assumption that rates h_y are normally distributed with mean μ_y and variance σ_y^2 but relax the assumption of independence by interval. Successive rates therefore have a non-zero covariance. Since the expected value of a sum is equal to the sum of expected values, the mean value of any period-specific pseudorate is still $\sum_t \mu_y$. However, the variance of the pseudorate is no longer equal to $\sum_t \sigma_y^2$. Moreover, the pseudorates are no longer normally distributed, and the distribution of year-end populations will not be log-normal.

We claim that the distribution of year-end populations exposed to random rates in a single-state population model will be given by the distribution of the exponential of some function $f(\cdot)$, where the arguments of $f(\cdot)$ are random rates not necessarily independently, normally or identically distributed. However, because of the nature of the population model, $f(\cdot)$ is in fact restricted to a simple sum, i.e. $f(h_y) = \sum h_y$. In the null model, attention was restricted to the case where arguments of $f(\cdot)$ are independent normal random variables. In the correlated case we drop the restriction of the independence of the h_y and consider non-independent normally distributed random variables. The sums of such random variables are not normally distributed, which has implications for calculating the moments of $e^{f(h)}$. We shall see that, without much additional complication, we can subsequently drop the assumption of normality. In all of these cases, our aim is to find expressions for the distribution of $e^{f(h)}$, or at least for its mean and variance. In the null model, we can determine the distribution of $e^{f(h)}$ exactly; in other cases we can do so only approximately.

A natural approach is to characterize the distribution of $e^{f(h)}$ in terms of the mean and variance of $f(\cdot)$. Such an approach is attractive since $f(\cdot)$ is always additive and its mean and variance can always be easily obtained. Although the mean will always be given by the sum of the means of its arguments, to obtain the variance of $f(\cdot)$, the covariance between rates needs to be taken into account. For example, the variance of the sum of two random rates, h_1 and h_2 , is given by:

$$\text{Eq. (17)} \quad \text{Var}(f) = \text{Var}(h_1 + h_2) = \text{Var}(h_1) + \text{Var}(h_2) + 2\text{Cov}(h_1, h_2).$$

For a pseudorate in year 3, three possibly non-zero covariance terms need to be considered:

$$\begin{aligned} \text{Eq. (18)} \quad \text{Var}(h_1 + h_2 + h_3) &= \text{Var}(h_1) + \text{Var}(h_2) + \text{Var}(h_3) + \\ &2\text{Cov}(h_1, h_2) + 2\text{Cov}(h_1, h_3) + 2\text{Cov}(h_2, h_3) \end{aligned}$$

It follows that, for the variance of the pseudorate in year t , there are $\binom{t}{2}$ potentially non-zero covariance terms.

Despite the growing number of terms in the variance of $f(\cdot)$, provided that the means, variances and covariances of the h_y are known, the mean and variance of $f(\cdot)$ can be exactly determined. A simple model for non-independent rates that satisfies this condition is one where rates follow an AR-1 process. Under such or similar conditions, therefore, determining the exact mean and variance of $f(\cdot)$ is feasible.

Series expansion methods

Approximating the mean and variance of $e^{f(h)}$ can be done with methods of series expansion. If it is sufficient to estimate the mean and variance of $e^{f(h)}$, one approach is to express $e^{f(h)}$ as a series expansion in terms of powers of a dummy constant and to take the required moments of the terms of a truncated version of the series (Weigel, 1986). If, on the other hand, one needs to characterize more fully the distribution of $e^{f(h)}$, the methods of Edgeworth expansion, whereby the density of $e^{f(h)}$ is approximated in terms of a modified normal density, are available (Kolassa, 2006).

Taking the first approach, we write $e^{\lambda f}$ as a series expansion in the powers of λ and take the mean of both sides (equivalent to writing f in terms of its characteristic function):

$$\text{Eq. (19)} \quad \langle e^{\lambda f} \rangle = 1 + \lambda \langle f \rangle + \frac{1}{2} \lambda^2 \langle f^2 \rangle + \dots$$

The right-hand side can then be rewritten (Weigel, 1986):

$$\text{Eq. (20)} \quad \langle e^{\lambda f} \rangle = \exp\{\langle \lambda f \rangle + \frac{1}{2} \lambda^2 (\langle f^2 \rangle - \langle f \rangle^2) + O(\lambda^3 f^3)\}.$$

Setting $\lambda = 1$, the mean of $e^{f(h)}$ can be written in terms of the mean, μ , and variance, σ^2 , of $f(\cdot)$ and a term that is $O(f^3)$:

$$\text{Eq. (21)} \quad \langle e^f \rangle = \exp\{\mu + \frac{1}{2} \sigma^2 + O(f^3)\}.$$

Note that when f is normally distributed, $O(f^3)$ must be identically equal to zero.

Truncating Eq. (21) at second order in the exponential, the mean of $e^{f(h)}$ can thus be written as approximately:

$$\text{Eq. (22)} \quad \langle e^f \rangle \approx \exp\{\langle f \rangle - \frac{1}{2} (\langle f^2 \rangle - \langle f \rangle^2)\}.$$

To obtain the variance of $e^{f(h)}$, we proceed as follows. Since:

$$\text{Eq. (23)} \quad \text{Var}(X) = \langle X^2 \rangle - \langle X \rangle^2,$$

it follows that:

$$\text{Eq. (24)} \quad \text{Var}(e^f) = \langle e^{2f} \rangle - \langle e^f \rangle^2.$$

But since, by Eq. (21):

$$\text{Eq. (25)} \quad \langle e^{2f} \rangle = \exp\{2\langle f \rangle - \frac{1}{2} (\langle 4f^2 \rangle - \langle 2f \rangle^2) + O(f^3)\},$$

and:

$$\text{Eq. (26)} \quad \langle e^f \rangle^2 = \exp\{2\langle f \rangle + \langle f^2 \rangle - \langle f \rangle^2 + O(f^3)\},$$

it follows that:

$$\text{Eq. (27)} \quad \text{Var}(e^f) \approx \exp\{2\langle f \rangle - 2\langle f^2 \rangle + \langle f \rangle^2\} - \exp\{2\langle f \rangle + \langle f^2 \rangle - \langle f \rangle^2\},$$

with exact equality when f is normal. So approximate expressions for the mean and the variance of $e^{f(h)}$ can be written in terms of the mean and variance of f .

The above approach is equivalent to approximating the exponential of an arbitrary random variable by the exponential of a normally distributed random variable. This is evidently analogous to approximating an arbitrary distribution by a normal distribution with the same mean and variance. While this may seem to be a somewhat arbitrary procedure, it is equivalent to truncating the Edgeworth expansion of f at the first term (Kolassa, 2006);

an expression for the error of truncation is moreover obtained. When f is even approximately normal, this is a reasonable procedure and is in fact commonly used (Weigel, 1986).

However, it is straightforward to obtain a better approximation if necessary by writing the distribution of $e^{f(h)}$ directly in terms of its Edgeworth expansion (also referred to as the cumulant expansion):

$$\text{Eq. (28)} \quad e^{f(h)} \sim \phi(e^f) \times [1 + \kappa_3 h_3(e^f)/6 + (\kappa_4 h_4(e^f)/24 + 10\kappa_3^2 h_6(e^f)/720) + \dots],$$

where $\phi(e^f)$ is the normal density function with the same mean and variance as e^f , the κ_i are the cumulants of e^f of order i and $h_i(e^f)$ are the Hermite polynomials (McCullagh, 1987). The higher-order cumulants of e^f ($i > 2$) will in general have to be estimated from a sample (possibly obtained by simulation) of e^f . The sample-dependent nature of the estimates of the higher-order cumulants makes it relevant to express the error in truncation in terms of a measure of sample size. However, even for theoretical calculations, it is almost never necessary to include more than the first or second non-constant term in square brackets (McCullagh, 1987). Truncation error can be expressed in terms of the sample size, n , used for estimation of the cumulants. Truncating the series after the first non-constant term in square brackets yields error $O(n^{-1})$; truncating after the second term (i.e. not including the dots) yields error $O(n^{-3/2})$ (McCullagh, 1987).

Recall that the covariance of populations depends only on the mean value of year-end populations (e.g. in the independent case $\alpha_y = e^{-\mu_y + \frac{1}{2}\sigma_y^2}$). In the dependent case, therefore, α_y will depend only on the mean of $e^{f(h)}$. Thus dependence in rates affects the correlation of populations only indirectly, in other words, through the means of the pseudorates. In the dependent case as well, the pseudorates therefore remain the key to the uncertainty analysis of the population model.

Implications of various random models for h

It is possible to make several qualitative statements about the behaviour of $f(\cdot)$ for different random distributions h :

- When h is uniformly distributed, Σh rapidly converges to an approximately normal distribution as the number of terms h increases.

- When h is gamma distributed with a common parameter θ , Σh will be gamma distributed, so when h is gamma distributed with parameters θ_y approximately equal, Σh will be approximately gamma distributed.
- When h is log-normally distributed, Σh will be approximately log-normally distributed (Wu et al., 2005).

Population model under uncertainty

The conclusions of the cohort model generalize relatively easily to the full population model. Considering other age-group cohorts, the conclusions stated above remain valid: under the null model, year-end populations are log-normally distributed random variables whose means and variances follow a simple additive-in-the-exponent rule and null-model populations can be characterized in terms of stochastic processes whose covariance properties are well established.

One additional complication is possible, however. Rates may be either correlated or uncorrelated across age groups. However, whatever the correlation pattern across age-specific rates, age-specific populations across the respective cohorts will be correlated through their respective population means.

Multi-state life table under uncertainty

Introduction

When there is more than one possible state in the population model, as in the longitudinal population model described in (Lauer et al., 2003), we have the exponential of the matrix of transition rates (Eq. (2)) rather than the exponential of a single transition rate. The exponential of a matrix is defined in terms of the infinite series of the powers of the matrix (cf. Singer and Spilerman, 1976). However, for diagonalizable matrices (i.e. for square matrices of full rank) the matrix exponential has a closed-form expression in terms of the eigenvalues and eigenvectors. A diagonalizable matrix C can be written

$$\text{Eq. (29)} \quad C = N \cdot \Lambda \cdot N^{-1},$$

where N is the matrix whose columns are the eigenvectors and Λ is a diagonal matrix of the eigenvalues. It is then easy to show that

$$\text{Eq. (30)} \quad \exp(\mathbf{C}) = \mathbf{N} \cdot \exp(\mathbf{\Lambda}) \cdot \mathbf{N}^{-1},$$

and that $\exp(\mathbf{\Lambda})$ is equal to the diagonal matrix whose elements are the exponentials of the individual eigenvalues (Strang, 1988). All $\mathbf{C}$ matrices for a population model will be square matrices of full rank, so these results apply to our case. In the population-model context, the eigenvalues are analogous to the rate in the scalar case; the eigenvectors represent the initial population values, and it is seen in the equation above, as in the equation for the scalar case, that they multiply the exponentiated eigenvalues in the matrix exponential (Caswell, 2001). Each distinct eigenvalue of $\mathbf{C}$ is associated with a unique state in the population model, and expresses the “long-term growth potential” of that state (Caswell, 2001).

When rates have units “year⁻¹”, the matrix exponential gives a matrix of probabilities whose i,j -th element expresses the one-year probability of transition from state i to state j in the population model (Singer and Spilerman, 1976). As such, each element in the matrix exponential represents the sum of the probabilities of making the $i \rightarrow j$ -transition considering all possible paths from state i to state j , including paths requiring an infinite number of steps (Van Kampen, 1992). As it affords what is called a “spectral decomposition” of the transition matrix $\mathbf{C}$, the matrix exponential provides a very useful formalism for the multi-state population model. The matrix exponential representation allows an analogous representation to that given for the single-cohort case in Eq. (3):

$$\text{Eq. (31)} \quad \mathbf{p}(t) = \mathbf{A}(\mathbf{C}) \cdot \mathbf{p}(t-1).$$

Here $\mathbf{A}(\mathbf{C})$ is the notation adopted for the matrix of survival probabilities (written as a function of the transition operator $\mathbf{C}$ to emphasize its dependence on the rates).

When rates are random, $\mathbf{A}(\mathbf{C})$ can be partitioned, as in Eq. (11), into a matrix of the constant means and a matrix of multiplicative random deviations. The multi-state longitudinal population model can then be written:

$$\text{Eq. (32)} \quad \mathbf{p}(t) = \langle \mathbf{A}(\mathbf{C}) \rangle \otimes \mathbf{d}\langle \mathbf{A}(\mathbf{C}) \rangle \cdot \mathbf{p}(t-1),$$

where $\langle \mathbf{A}(\mathbf{C}) \rangle$ refers to the matrix of means of the elements of $\mathbf{A}(\mathbf{C})$, $\mathbf{d}\langle \mathbf{A}(\mathbf{C}) \rangle$ refers to the matrix of multiplicative random deviations around the means and " $\otimes$ " indicates the Hadamard, or element-by-element, product.

In analogy, moreover, with the constant factors α_j of Eq. (13), the elements of $\langle \mathbf{A}(\mathbf{C}) \rangle$ are expressable in terms of the series expansion of the elements of $\mathbf{A}(\mathbf{C})$. For example, if an element of $\mathbf{A}(\mathbf{C})$, which we can notionally represent as e^f (the notion being that f is an eigenvalue of $\mathbf{C}$), happens to be log-normally distributed, its mean as given by Eq. (21) will not contain a term of $O(f^3)$ but will be exactly equal to $e^{\mu + \frac{1}{2}\sigma^2}$, where μ and σ^2 are the mean and variance, respectively, of f (Gardiner, 2004). In this case, the corresponding element in the matrix of deviations, $\mathbf{d}\langle \mathbf{A}(\mathbf{C}) \rangle$, is a log-normally distributed random variable with a mean of 1 and variance equal to $e^{-2\mu + \sigma^2}(-1 + e^{\sigma^2})$.

Distribution of elements of $\mathbf{A}(\mathbf{C})$

Since, however, each element of $\mathbf{A}(\mathbf{C})$ is in fact of the form $\sum_i g_i \cdot e^f g_i^{-1}$, where g_i represents eigenvector i of $\mathbf{C}$ and f represents an eigenvalue of $\mathbf{C}$, elements of $\mathbf{A}(\mathbf{C})$ will in general not be log-normal in distribution. However, algebraic expressions for the eigenvalues and eigenvectors of the matrix $\mathbf{C}$ can be calculated for any arbitrary population model with four or fewer states. In addition, for many reasonable population models containing five or more states, the eigenvectors and eigenvalues can still be calculated as algebraic expressions. Entries in $\mathbf{A}(\mathbf{C})$ can therefore often be expressed in terms of analytic functions of algebraic expressions of random variables.

For example, for a population model with the 2×2 matrix $\mathbf{C}$:

$$\begin{pmatrix} a + c & -b \\ -c & a + b \end{pmatrix},$$

the eigenvalues are:

$$\begin{pmatrix} a \\ a + b + c \end{pmatrix},$$

the eigenvectors are the columns of:

$$\begin{pmatrix} b/c & -1 \\ 1 & 1 \end{pmatrix},$$

and the corresponding matrix $\mathbf{A}(\mathbf{C})$ is:

$$\begin{pmatrix} \frac{e^a (b + ce^{b+c})}{b+c} & -\frac{be^a (-1 + e^{b+c})}{b+c} \\ -\frac{ce^a (-1 + e^{b+c})}{b+c} & \frac{e^a (c + be^{b+c})}{b+c} \end{pmatrix}.$$

What can we say in this case about the distribution of the elements of $\mathbf{A}(\mathbf{C})$ when the elements of $\mathbf{C}$ are random? If one assumes that rates are normally distributed and independent, their sums and differences will also be normally distributed. In addition, the exponentials of normally distributed random variables are log-normally distributed. So the numerators of the individual elements can be reduced to a sum of the products of normal and log-normal distributions. Since the denominators in this case are themselves normal distributions, it is reasonable to expect that the ratio distribution will be approximately log-normally distributed (as an heuristic, one can imagine that the normal densities “cancel” in both the numerator and denominator).

However, for a population model such as that in (Lauer et al., 2003), with a 4×4 matrix $\mathbf{C}$, these expressions contain hundreds of terms, including polynomial combinations of elements of $\mathbf{C}$ and roots of polynomial combinations of elements of $\mathbf{C}$. In such a case, no obvious intuition seems available concerning the distribution of the elements of $\mathbf{A}(\mathbf{C})$; calculating the distributional form analytically is moreover a virtual impossibility.

We suggest the following conjecture. Since, when many random terms combine additively a normal distribution is the result, and when many such terms combine multiplicatively a log-normal distribution results, when many random terms combine both additively and multiplicatively, depending on the predominance of one or the other operation, as well as on the relative magnitudes of the terms, one can expect either the normal or the log-normal

form to dominate, at least as an approximation. We support this conjecture below with results from simulations.

Cumulant expansion of elements of $\mathbf{A}(\mathbf{C})$

As both the eigenvalues (f) and eigenvector elements (g) of $\mathbf{C}$ are functions of random variables, the cumulant expansion of the elements will depend on the distributions of g and of f , as well as on their correlations. Moreover, as there is no guarantee that elements of $\mathbf{A}(\mathbf{C})$ are normal in distribution, their cumulant expansion will in general have a potentially infinite number of terms. Consequently explicit expressions for their means and variances will necessarily be approximate, and the elements of $\mathbf{d}\langle\mathbf{A}(\mathbf{C})\rangle$ can likewise be written as only approximately equal to 1 in expectation. Although this implies that we are limited to writing Eq. (26) as an approximate equality only, any arbitrary level of precision can be obtained by means of additional terms in the cumulant expansion if necessary.

Simulation approach

This difficulty of calculating an explicit distributional form for the elements of $\mathbf{A}(\mathbf{C})$ when its elements are not normally distributed suggests an efficient use of the simulation method for purposes of studying the uncertainty of a given multi-state longitudinal population model:

1. One defines distributions for the random rates that make up the elements of $\mathbf{C}$.
2. Multiple random draws are made from each of these distributions, so as to produce a distribution of random matrices $\mathbf{C}$.
3. For each draw of $\mathbf{C}$, the matrix of one-year probabilities $\mathbf{A}_y(\mathbf{C}_y)$, which is in general understood to be period specific, is computed numerically.
4. The resulting distribution of random matrices $\mathbf{A}_y(\mathbf{C}_y)$ is used to compile distributions for each of its elements.
5. The distributions of each of the elements of $\mathbf{A}_y(\mathbf{C}_y)$ is used to calculate numerically the desired number of cumulants for each of the element distributions.
6. Using the results from step (5), the matrices $\langle\mathbf{A}(\mathbf{C})\rangle$ and $\mathbf{d}\langle\mathbf{A}(\mathbf{C})\rangle$ are constructed.
7. The population model is then approximately represented in the form of Eq. (26).

8. When the uncertainty of multiple-year life-table quantities are required, the matrix products $\Pi_t \langle \mathbf{A}_y(\mathbf{C}_y) \rangle$ and $\Pi_t \mathbf{d} \langle \mathbf{A}_y(\mathbf{C}_y) \rangle$ are calculated.
9. The product $\Pi_t \langle \mathbf{A}_y(\mathbf{C}_y) \rangle$ gives the correlation constant for the population model and the product $\Pi_t \mathbf{d} \langle \mathbf{A}_y(\mathbf{C}_y) \rangle$ gives the random component.
10. The delta method can be used to calculate the mean and variance of derived life-table quantities like population residence time.

It can thus be seen that the distribution of the elements of the matrix $\mathbf{A}_y(\mathbf{C}_y)$ is sufficient for defining the uncertainty characteristics of any life-table quantity in the multi-state population model. Just as for the pseudorates, the exponential component of elements of $\Pi_t \mathbf{A}_y(\mathbf{C}_y)$ will show additivity in the eigenvalues, so it may be possible to achieve some simplifications in the above methods (e.g. calculating only the distributions of the eigenvalues of $\mathbf{C}$). However, for the purposes of simulation it will usually be just as convenient to calculate the full expression for the components of $\Pi_t \mathbf{A}_y(\mathbf{C}_y)$.

Simulation results

Regarding our conjecture concerning the distributional form of the elements of $\mathbf{A}_y(\mathbf{C}_y)$, we stated above that it should be either approximately normal or approximately log-normal. This is indeed what is found in simulations based on population models such as the one presented in (Lauer et al., 2003). In fact, these simulations suggest that elements of $\mathbf{A}(\mathbf{C})$ show a strong tendency overall to converge towards an approximately normal distribution (Figure). The histograms of the elements of $\mathbf{A}_y(\mathbf{C}_y)$ display a bell-shaped curve well fit by the normal density, that shows either negligible skewness or a slight log-normal skewness).

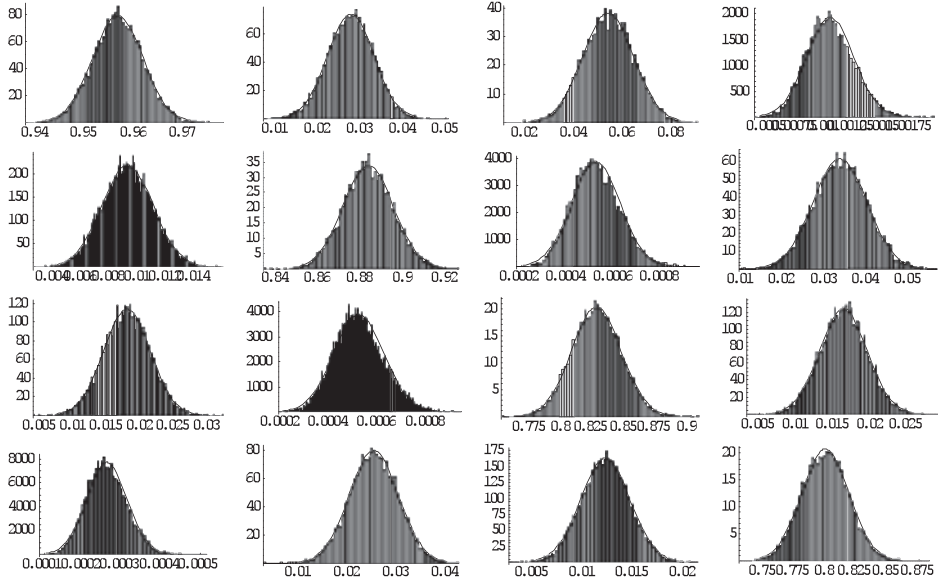


Figure: Histograms of elements of $\mathbf{A}(\mathbf{C})$ for the population model in [Lauer et al., 2003], showing the corresponding line of normal density. The distributions were obtained from 10,000 simulations where the elements of $\mathbf{C}$ are independent and normally distributed.

Notice, however, that the distribution of anti-diagonal elements $\{(i, j): (1, 4), (2, 3), (3, 2), (4, 1)\}$ in the array are noticeably more skew (i.e. more log-normal) than the distribution of other elements, which are only negligibly if at all skew. This pattern has a meaning in terms of the population model in [Lauer et al., 2003], since the anti-diagonal elements represent transitions that cannot be made in a single step, that is $\{(j \rightarrow i): (4 \rightarrow 1), (3 \rightarrow 2), (2 \rightarrow 3), (1 \rightarrow 4)\}$, but only as a result of a series of two or more steps.

Thus, the shortest path that implements the transition $\{(4 \rightarrow 1) = (XC \rightarrow S)\}$ is either the path $\{(XC \rightarrow C) \cup \{(C \rightarrow S)\}$ or the path $\{(XC \rightarrow X) \cup \{(X \rightarrow S)\}$. Similarly, the shortest path implementing the transition $\{(3 \rightarrow 2) = (X \rightarrow C)\}$ is either the path $\{(X \rightarrow XC) \cup \{(XC \rightarrow C)\}$ or the path $\{(X \rightarrow S) \cup \{(S \rightarrow C)\}$. Since the probabilities of making a double or higher-order transition always involve the product of at least two probabilities, it is reasonable to expect that the distributions of the probabilities of these transitions should be more log-normal.

The same is true of the diagonal probabilities, which represent the probability of starting in a state and ending in that same state in the same interval. The relevant paths also contain at least two steps. Taking the logarithm of the distributions from the simulations shows that, for all distributions other than the diagonal and anti-diagonal distributions, the logarithmically transformed distributions are distorted whereas the transformed diagonal and anti-diagonal distributions appear symmetric. This suggests that the diagonal and anti-diagonal distributions have a more pronounced tendency to converge to log-normal distributions than the other distributions.

These findings suggest the possibility of making the following emendation to the simulation procedure, above. After step (4), insert:

- 4a. If the distributions of any of the elements of $\mathbf{A}_j(\mathbf{C}_j)$ are noticeably skew, take the logarithm of the values and calculate the cumulants of this transformed distribution (where a low-order truncation of the cumulant expansion will be more justified). Exponentiate to recover the original distribution.

Conclusions

The problem of determining the distributional features of $\mathbf{p}(t)$ in a multi-state longitudinal population model can be seen to reduce to the problem of determining the distribution of elements of $\mathbf{A}(\mathbf{C})$ when $\mathbf{C}$ is a random matrix. The demographic or statistical literature on Markov models does not appear to address the question of random elements in the operator matrix $\mathbf{C}$. The indicative results we present here are nevertheless a straightforward extension of the linear algebra theory that underlies most multi-state modelling theory.

References

1. Caswell H. *Matrix population models: construction, analysis, and interpretation*. Sunderland, Massachusetts, Sinauer Associates, 2001.
2. Franke J, Härdle W, Hafner C. *Statistics of Financial Markets*. Berlin, Springer, 2004.
3. Gardiner CW. *Handbook of stochastic methods for physics, chemistry and the natural sciences*, 3rd ed. Berlin, Springer, 2004.
4. Goldberger AS. *A course in econometrics*. Cambridge, Massachusetts, Harvard University Press, 1991.
5. Kolassa JE. *Series Approximation Methods in Statistics*, 3rd ed. Springer, 2006.
6. Lauer JA, Roehrich K, Wirth H, Charette C, Gribble S, Murray CJL. PopMod: a longitudinal population model with two interacting disease states. *Cost effectiveness and resource allocation*, 2003, 1:6.
7. McCullagh P. *Tensor methods in statistics*, London, Chapman and Hall, 1987.
8. Singer B, Spilerman S. The representation of social processes by Markov models. *The American journal of sociology*, 82(1):1-54, 1976.
9. Gilbert Strang. *Linear algebra and its applications*, 3rd ed. Harcourt Brace Jovanovich, 1988.
10. Van Kampen NG. *Stochastic processes in physics and chemistry*. Amsterdam Elsevier, 1992.
11. Weigel FW. *Introduction to Path-Integral Methods in Physics and Polymer Science*, Singapore, World Scientific Publishing Company, 1986.
12. Wu J, Mehta NB, Zhang J. Flexible Lognormal Sum Approximation Method, IEEE Global Telecommunications Conference (GLOBECOM), Vol. 6, pp. 3413-3417, November 2005 (IEEE Xplore).