Mortality: a statistical approach to detect model misspecification

Jean-Charles Croix, Frédéric Planchet, Pierre-Emmanuel Thérond

Bulletin Français d'Actuariat manuscript No.
(will be inserted by the editor)

Mortality: a statistical approach to detect model misspecification

Jean-Charles Croix · Frédéric Planchet · Pierre-E. Thérond

April 30th, 2015

Abstract The Solvency 2 advent and the best-estimate methodology in future cash-flows valuation lead insurers to focus particularly on their assumptions. In mortality, hypothesis are critical as insurers use best-estimate laws instead of standard mortality tables. Backtesting methods, i.e. ex-post modeling validation processes, are encouraged by regulators and rise an increasing interest among practitioners and academics. In this paper, we propose a statistical approach (both parametric and non-parametric models compliant) for mortality laws backtesting under model risk. Afterwards, a specification risk is introduced assuming that the mortality law is subject to random variations. Finally, the suitability of the proposed method will be assessed within this framework.

Keywords Solvency 2 · mortality · cusum · detection · SPRT

1 Introduction

The Solvency 2 directive (art. 83, Comparison against experience) imposes that undertakings develop processes to ensure that Best-Estimate calculations and underlying hypotheses are regularly compared against experience. In Life insurance and particularly in annuity computations, mortality models validation and backtesting is of key importance.

In this context, we consider the following simple question: How does an insurer

Jean-Charles Croix · Frédéric Planchet · Pierre-E. Thérond
ISFA - Université Lyon 1, Laboratoire SAF, 50 avenue Tony Garnier, 69007 Lyon, France

Jean-Charles Croix
E-mail: jean.charles.croix@gmail.com

Frédéric Planchet
Prim'Act - 42 avenue de la Grande Armée - 75017 Paris
E-mail: frederic@planchet.net

Pierre-E. Thérond
Galea & Associés - 12 avenue du Maine - 75015 Paris - France
E-mail: ptherond@galea-associes.eu & pierre@therond.fr

verify that his mortality hypotheses are Best-Estimate ? More precisely, we derive testing methodologies to decide whether a given table is likely, according to observations. Indeed, the insurer wants to distinguish sampling variations from misspecification at any age. To do so, a reminder of mortality analysis and models is provided in a first part. The derived statistical models are adequate foundations to develop and support testing processes that detects if prediction errors are the result of sampling variations or an unknown trend. According to these models, a first set of tests with fixed sample sizes are reviewed.

In a second part, the review will be extended to on-line backtesting, which relies on tests with random sample sizes. Indeed, if an insurer repeats fixed tests on a growing set of data (every month for example), the first type error probability converges to one if no corrections are taken on the significance level. This problem is solved using sequential analysis and change-point detection algorithms. Finally, a numerical application is proposed to compare different approaches faced to a simulated misspecification.

2 Mortality models & assumptions

In mortality analysis, life time is considered as a positive random variable T. Considering sufficiently large groups of individuals, mortality risk is assumed mutualized and mathematical models are employed to describe the average behavior of a specific population. Writing S and h the survival and hazard functions respectively, the probability of death between age x and $x+1$ (i.e. at age x) can be expressed as in equation 1 (see Planchet and Thérond (2011)):

$$q_x = P(T \leq x+1 | T > x) = 1 - \frac{S(x+1)}{S(x)} = 1 - \exp\left(-\int_x^{x+1} h(u)du\right). \quad (1)$$

If one wants to predict the number of deaths in a population for a fixed period (without any other causes of population reduction), a minimal segmentation is needed to obtain homogeneity: a simple classifier is age. Under these assumptions, the number of deaths D_x at age x among a population of n_x individuals can be modelled as a binomial random variable. In a portfolio with p different ages $x \in [x_1, x_p]$, it comes:

$$\forall x \in [x_1, x_p],\ D_x \sim \mathcal{B}(n_x, q_x), \quad (2)$$

in case of annual projections. In the latter, mortality modeling will be summarized in an annual mortality table $q = (q_{x_1}, ..., q_{x_p})$. Furthermore, we will consider observations in monthly requiring a mortality table transformation. If death rates are supposed constant during one year, monthly mortality rates can be derived as follows:

$$_mq_x = 1 - {}_mp_x = 1 - (1 - q_x)^{\frac{1}{12}}, \quad (3)$$

where $_mq_x$ being the desired rate. In the following, all mortality rates are monthly, and the subscript m is omitted. This simple assumption can be refined according to the mortality model implied in table generation. A second assumption in this work is that population renew identically every time-step during analysis.

As a convention in this document, single letters designate vectors over ages (for example, the previously defined q represent a set of p death probabilities), the subscript x is age-specific (q_x is a real) and the upper-script represents different tables ($q^0 = (q^0_{x_1}, ..., q^0_{x_p})$ is the table underlying observations for example).

From a statistical view, and whichever the method used to produce the table, it can be considered as a parameter in a parametric model ($\mathcal{Y}$,$\mathcal{P}_Q$) with $\mathcal{Y}$ the set of all possible observations and $\mathcal{P}_Q$ a family of probability distribution on $\mathcal{Y}$ (see Gourieroux and Monfort (1996) for detailed developments and notations). All previous assumptions can be summarized in the following model:

$$\mathcal{M}_B = \left(\forall x \in [x_1, x_p], \mathcal{Y} = \mathbb{N}, \mathcal{P}_Q = \mathcal{B}\left(n_x, q_x\right), q_x \in Q_x\right), \tag{4}$$

with $\forall x \in [x_1, x_p], Q_x = [0, 1]$. If this model is well defined, and portfolio sizes are usually large, a Gaussian approximation is often used to simplify computations based on the central limit theorem. Even though this result is asymptotic (i.e. for large n_x), it's commonly used as the Gaussian law provides ease at use. Furthermore, we'll consider a fixed and known variance-covariance matrix, essentially for simplicity. Finally, we consider the following statistical model:

$$\mathcal{M}_G = \left(\forall x \in [x_1, x_p], \mathcal{Y} = \mathbb{R}, \mathcal{P}_Q = \mathcal{N}\left(\mu_x, \sigma_x^2\right), q_x \in Q_x\right), \tag{5}$$

with $\forall x \in [x_1, x_p]$, $\mu_x = n_x q_x$ and $\sigma_x^2 = n_x q_x (1 - q_x)$ (in vectorial notations $\mu = \left(\mu_{x_1}, ..., \mu_{x_p}\right)$ and $\left(\Sigma\right)_x = \sigma_x^2$ a diagonal matrix). From now on, we consider that we observe at each date i, a set of deaths $d^i = \left(d^i_{x_1}, ..., d^i_{x_p}\right)$ from which we build the gross mortality rates based on N observations, $\hat{q} = \left(\hat{q}_{x_1}, ..., \hat{q}_{x_p}\right)$ where $\forall x \in [x_1, x_p], \hat{q}^N_x = \frac{1}{N} \frac{\sum_{i=1}^N d^i_x}{n_x}$ (which is the maximum likelihood estimator in our model). Now that our framework for mortality models is defined, we shall present what our backtesting procedure is.

3 Mortality backtesting

Backtesting can be defined as an ex-post model validation method, including two different practices: validation and monitoring. The first aims to validate a mortality table according to a fixed amount of data, while the second allows for continuous treatment. This last aspect can be used to increase power in validation or detect shifts later on.

These problems are usually addressed through decision theory (see Gourieroux and Monfort (1996) or Saporta (2006) for detailed introductions). In our framework, it consists in testing the mean of a Gaussian vector with known variance and detecting any change-point or misspecification. One can find alternative approaches based on different setups (see El Karoui et al. (2013) for cox-like models and homogeneous Poisson processes).

Writing q^γ the supposed mortality table and q^0 the real one (i.e. which generates the data), the null hypothesis is $H_0 = \{q^\gamma = q^0\}$ against a composite alternative

$H_1 = \{q^\gamma \neq q^0\}$. Then tests are defined as couples (ξ_N, N) with N the sample size (possibly random) and ξ_N the associated decision function. All presented procedures are based on likelihood functions, derived from model $\mathcal{M}_G$ but all classical significance tests are applicable. Numerous other tests and change-point procedures can be found elsewhere, especially in change-point detection where the research is still very active in both Frequentist and Bayesian paradigms (see Lai (2001), Tartakovsky and Moustakides (2010b) and Tartakovsky and Moustakides (2010a)). On the other side, sequential alternatives are described in Wald (1947), Ghosh and Sen (1991), Siegmund (1985) and Basseville and Nikiforov (1993).

3.1 Fixed sample tests

Based on the above discussion, we consider fixed sample size tests in this section. In particular, Wald, Score and Likelihood ratio are easily applicable to the previous model and their asymptotic properties (convergence and coverage) are of importance as undertakings usually possess large portfolios. Following Gourieroux and Monfort (1996), we consider the multidimensional constraint $g(q) = q - q^\gamma$ which resume the simple hypothesis H_0. In the case of testing the mean of a Gaussian vector, these three tests correspond to the following statistic:

$$\xi = N(\hat{q}^N - q^\gamma)^T \Sigma^{-1} (\hat{q}^N - q^\gamma), \tag{6}$$

which is χ^2-distributed under H_0. The associated rejection region W is:

$$W = \{\xi > \chi^2_{1-\alpha}(p)\}, \tag{7}$$

p being the number of ages considered in the portfolio and $\chi^2_{1-\alpha}(p)$ the chi-square quantile with p-degrees of freedom and $1-\alpha$ level. By construction, fixed sample size tests require predefined parameters: a significance level α (or first term error probability) and a predefined sample size N (equivalent to time for periodic observations). In practice, insurers have to define when the test will be conducted: immediately or later with more information ? This decision implies a trade off between fast reaction and power: statistical significance increases with observation as mortality risk. Alternative tests can be found, based on the Standardized Mortality Ratio for example, see Liddell (1984) for example.

3.2 On-line backtesting

In this part, dynamic methods are investigated. The two main related theories are sequential analysis (see Wald (1947)) and change-point detection (Lai (2001) and Basseville and Nikiforov (1993) for detailed presentations and Tartakovsky and Moustakides (2010b) for a more recent review on bayesian technics). Indeed, a simple repetition of previous fixed sample size tests leads to important first type error probability increases. In the following, α is the probability to reject the null hypothesis when it's true and β the probability to keep the null hypothesis when the alternative is true (respectively the first and second type errors).

3.2.1 Sequential Probability Ratio Test (SPRT)

The sequential probability ratio test (SPRT) was first introduced as a test between two simple hypotheses. Constructed on the likelihood ratio Λ_N with N-observations:

$$\Lambda_N(x) = \prod_{i=1}^{N} \frac{\mathcal{L}(D = d^i, q^\gamma)}{\mathcal{L}(D = d^i, q^0)} \tag{8}$$

where p_n is the joint probability (or density) of the sample, the Sequential Probability Ratio Test consists in the following (with A,B two thresholds):

$$\begin{cases} reject\, H_0,\ if \Lambda_N \geq A \\ accept\, H_0,\ if \Lambda_N \leq B \\ continue,\ otherwise \end{cases} \tag{9}$$

In other terms, the test stops the first time the likelihood ratio leaves the interval $[B, A]$. The corresponding number of observations is called the sample size N and is thus a random variable. Optimality and closure properties are discussed in Wald (1947). Furthermore, the following approximations for A, B, α and β holds (even if the independance assumption is dropped):

$$\begin{aligned} A &\simeq \frac{1-\beta}{\alpha}, \\ B &\simeq \frac{\beta}{1-\alpha}. \end{aligned} \tag{10}$$

These expressions are only approximate due to possible overshoot over boundaries (Λ_n is never equal to A or B exactly when the test stops). In case of composite hypotheses, the situation is much more complex and the initial Likelihood Ratio Λ_n must be adapted. Wald (1947) proposed two different solutions. The first is a weighted sequential probability ratio test (WSPRT), obtained specifying prior distribution functions under H_0 and H_1 for the parameter to be tested. Indeed, in the case of a simple hypothese against a composite one, Wald proposes to consider the parameter as a random variable itself (and thus consider a specific law, v such as $\int_{Q_1} v(s)ds = 1$). Thus, he requires that the second type error probability β is controlled:

$$\beta = \int_{Q_1} \beta(q) w(q) dq \tag{11}$$

which is combersome as we would prefer to have $\forall q \in Q_1, \beta(q) \leq \beta$ (and no methods have been found to insure that constraint). The likelihood ratio becomes (the tilde notation will be reserved to probability mixture):

$$\tilde{\Lambda}_N(x) = \frac{\prod_{i=1}^{N} \mathcal{L}(D = d^i, q^\gamma)}{\int_{Q_1} \prod_{i=1}^{N} \mathcal{L}(D = d^i, q) v(q) dq} \tag{12}$$

The second is based on the generalized sequential probability ratio test (GSPRT), using estimators (usually Maximum Likelihood estimators) in place of priors (the hat notation will be used for likelihood estimators):

$$\hat{\Lambda}_N(x) = \prod_{i=1}^{N} \frac{\mathcal{L}(D = d^i, q^\gamma)}{\mathcal{L}(D = d^i, \hat{q})} \tag{13}$$

According to Wald (1947), this last version is more difficult to study as the likelihood ratio is no longer a probability distribution (in particular, approximations on error probabilities are not applicable). More recently, Lai (1998) proposed a dynamic boundary for the GSPRT, considering estimators variability.

As mentioned before, the main difficulty in the WSPRT design is the choice of an appropriate prior for the parameter q on Q_1. An existing solution is the frequency functions method, based on the likelihood ratio of a sequence of statistics. Using Cox's factorization theorem (in annex), one can reduce composite hypotheses to simple ones using an invariance reduction principle (see Hall et al. (1965) for further developments and Jackson and Bradley (1961) for a detailed application of this theorem). Applying this method to the gaussian case, Jackson and Bradley (1961) derived χ^2 (and T^2, in case of unknown variance-covariance matrix) sequential probability ratio test, based on homonym statistics. From now on, we apply their result to the previous backtesting problem even though they considered alternative hypotheses of the form: $H_0 = \{\|q^\gamma - q^0\| \leq \lambda_0\}$ against $H_1 = \{\|q^\gamma - q^0\| \geq \lambda_1\}$ with $0 \leq \lambda_0 < \lambda_1$ implying an indifference region (Depending whether acceptance is needed, one can set $\lambda_0 = 0$, which is the case in the following):

$$\ln \Lambda_N^{\chi^2} = -N\frac{\lambda_1^2}{2} + \ln {}_0F_1\left(\frac{p}{2}, \frac{N\lambda_1^2\chi_N^2}{4}\right). \tag{14}$$

where ${}_0F_1$ is the generalized hyper-geometric function and $\chi_N^2 = N(\hat{q}-q^\gamma)^T\Sigma^{-1}(\hat{q}-q^\gamma)$ (for numerical evaluation, we use the *gsl* package). This result is the ratio between two non-central χ^2 distributions with p-degrees of freedom and respective non-centrality parameter λ_1^2. The choice of λ_1 should be motivated by the application, and in our setup we have decided to select $\lambda_1 = 10\%\|q^\gamma\|$. The choice of A and B is based on Wald's previous approximation which still holds in this case. Unfortunately, there are no practical results to compute the expected sample sizes in this case (but they're are available under i.i.d assumption). One simplification suggested in Jackson and Bradley (1961) is to compute every time step independent statistics using only innovations: the Wald's approximation will hold in despite of a potentially substantial loss of power. No results are available on the GSPRT, thus we will not use it.

3.2.2 Quickest detection algorithms

Backtesting can also be interpreted as a change-point detection problem. In this theory, the classical setup is a sequence of random variables distributed under a known distribution f_0, that possibly switches to an alternative and unknown distribution f_1 at an unknown time $\nu \in \mathbb{N}$ (random in Bayesian frameworks and considered equal to ∞ when no changes occur). The objective for change-point detection algorithm τ (defined as a stopping time) is to raise an alarm as quickly as possible when the change occurs, without raising too frequent false alarms. According to Tartakovsky, 4 approaches can be found in the literature: Bayesian (the time of change is random with a specific prior), Generalized Bayesian (improper priors), Multi-cycle procedures and Minimax. Change-point detection is a vast domain and we will focus on frequentist algorithms. Lorden (1971) gave a minimax criterion to compare algorithms in this setup, the essential supremum

average detection delay (Notations from Tartakovski, ν is the time of change, τ is the time where the alarm is raised):

$$ESADD = \sup_{0\leq\nu<\infty} \operatorname{ess\,sup} E_\nu \left[(\tau-\nu+1)^+ | \mathcal{F}_\nu\right]. \tag{15}$$

subject to a constraint of maximal false alarm frequency $E_0(\tau) \geq \lambda$. In other terms, we want our process to raised the alarm as quicly as possible when the change as occured, without raising a false alarm before time λ (in average) when there is no change ($\nu = 0$). As a solution to this problem, Page (1954) introduced the Cusum algorithm:

$$\tau = \inf\{n, \Lambda_n - \min_{1\leq j\leq n} \Lambda_j \geq A\} = \inf\{n, \max_{1\leq j\leq n} \Lambda_j^n \geq A\}, \tag{16}$$

where Λ_j^k being the likelihood ratio based on observations j up to k. A recursive version of this algorithm can be found in Lorden (1971) and Basseville and Nikiforov (1993). In his work, Page also pointed out the connection between the Cusum algorithm and Wald's SPRT: the Cusum test can be seen as a set of parallel open-ended SPRTs, a new one starting every period (or observation). Writing N_k the sample size of a one-sided open-ended SPRT applied to $\hat{q}_k, \hat{q}_{k+1}, ...$, the Cusum stopping time is $N^\star = \min_{1\leq k\leq n} N_k$.

Equivalently to the SPRT, two solutions are presented to deal with composite hypotheses:

- the Weighted Cusum $\tilde{\Lambda}$:

$$\tilde{\Lambda}_j^k = \int_{q^1\in\Theta} \frac{\mathcal{L}(\hat{q}^j, ..., \hat{q}^k | q^1)}{\mathcal{L}(\hat{q}^j, ..., \hat{q}^k | q^0)} dF(q^1), \tag{17}$$

- the Generalized Likelihood Ratio (GLR) $\hat{\Lambda}$:

$$\hat{\Lambda}_j^k = \frac{\sup_{q^1\in\Theta_1} \mathcal{L}(\hat{q}^j, ..., \hat{q}^k | q^1)}{\mathcal{L}(\hat{q}^j, ..., \hat{q}^k | q^0)}. \tag{18}$$

Considering previous alternative hypotheses and following Basseville and Nikiforov (1993), two χ^2-Cusum algorithms are available. The first is a direct application (case 3 p.218) of least favorable priors in case of invariant distributions:

$$\ln \tilde{\Lambda}_j^k = -(k-j+1)\frac{\lambda_1^2}{2} + \ln {}_0F_1\left[\frac{p}{2}, \frac{(k-j+1)\lambda_1^2(\chi_j^k)^2}{4}\right], \tag{19}$$

with $(\chi_j^k)^2 = (k-j+1)\left(\hat{q}_j^k - q^\gamma\right)^T \Sigma^{-1}\left(\hat{q}_j^k - q^\gamma\right)$. Asymptotic first-order optimality has been proven for the χ^2-Cusum algorithm in multidimensional case (see p.268 in Basseville and Nikiforov (1993)). Introducing $\Delta_0 = E_{\theta_0}(N^\star)$ the mean-time between false alarms and $\Delta_1 = E_{\theta_1}(N^\star)$ the average delay for detection, it comes:

$$\Delta_0 \geq A, \tag{20}$$

from Lorden's theorem (see theorem in annex). Furthermore, the χ^2-Cusum algorithm is first-order optimal in that case.

4 Numerical applications

In this section, we propose a simple numerical illustration to ensure tests efficiency in the context of mortality backtesting. Tests will first be tested under null hypothesis and then in case of mortality table misspecified. This case will be simulated in a practical method, using a white noise on mortality rates logits. After unbiasing, the table we consider as the real mortality tables q^0 is randomly distributed around the given mortality table q^γ, but equal in mean.

4.1 Misspecification on mortality tables

Specification risk occurs when the given mortality table doesn't fit the real mortality distribution. In this case, if q^0 is the real mortality law, q^γ the model and ϵ the error term it comes:

$$q^0 = f(q^\gamma, \epsilon), \tag{21}$$

where f is an unknown and unobservable function and ϵ a random variable. In this application, our methodology consists in choosing a specific function f and a probability distribution for the error term to produce specification risk. The error term is a controlled Gaussian white noise applied to the pre-defined mortality law logits:

$$\forall x \in [x_1, x_p],\ \text{logit}(q_x^0) = \text{logit}(q_x^\gamma) + \epsilon_x, \tag{22}$$

with $\epsilon \sim \mathcal{N}_p(0, \sigma Id)$. In other words, the real mortality law is randomly distributed around the pre-defined law q^γ but equal in average ($E(q^0) = q^\gamma$). Thus, the function f is the following:

$$\forall x \in [x_1, x_p],\ q_x^0 = \frac{e^{\epsilon_x} q_x^\gamma}{1 + q_x^\gamma (e^{\epsilon_x} - 1)} - E\left(\frac{e^{\epsilon_x} q_x^\gamma}{1 + q_x^\gamma (e^{\epsilon_x} - 1)} - q^\gamma\right). \tag{23}$$

Finally, an illustration is given of multiple q^0 randomly distributed around q^γ (see figure 1).
Now that specification risk is simulated, the second objective is to find a business interpretation of σ. Indeed, if it's quantitatively defined in previous equations, what impact does-it have on real indicators ? For instance, the volatility implied on the remaining life expectancy of a 65-years old male e_{65} is measured as follows:

$$e_{65} = \frac{1}{S(65)} \sum_{j=66}^{120} S(j), \tag{24}$$

with $S(x) = \prod_{i=1}^{x-1} (1 - q_i)$ the discrete survival function. Considering e_{65} as a function of ϵ, here is a measure of the deviation of e_{65}:

$$\delta = \frac{q_{95\%}(e_{65}) - E(e_{65})}{E(e_{65})}. \tag{25}$$

The following table 1 shows correspondence between remaining life expectancy volatility δ and σ.

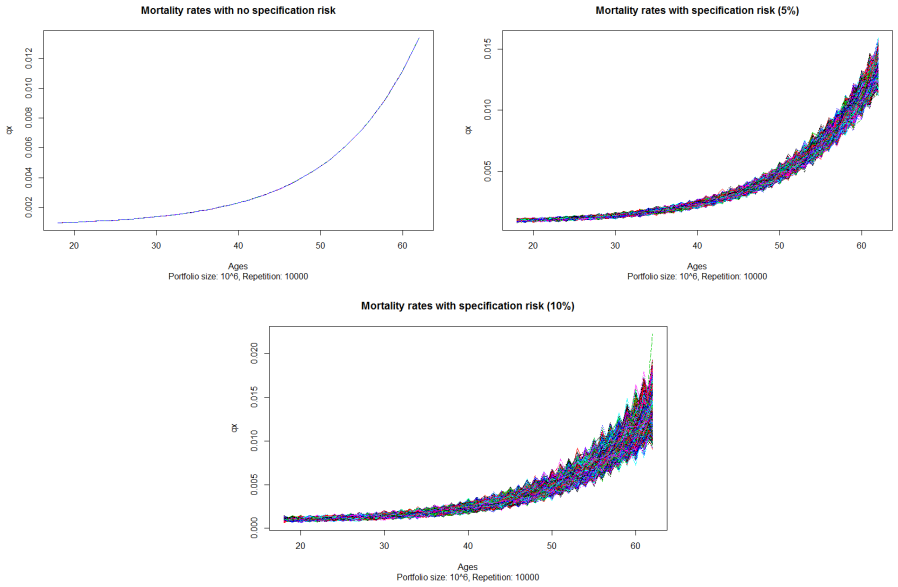

Fig. 1 Example of different levels of specification risk $(0, 5\%, 10\%)$.

Table 1 Correspondence between σ and δ for a 65 years old person and $N = 10^6$.

σ	e	δ
0%	16.21	0.00000
5%	16.34	0.00708
10%	16.48	0.01556
20%	16.75	0.03051
30%	17.00	0.04770
40%	17.23	0.06508

4.2 Data simulation and portfolio structure

The test methodology consists in setting first q^γ (in our example, it has been adjusted on the French regulatory mortality table TH00-02). Then, for each simulation, a noise is simulated and applied to obtain q^0. From that, deaths are generated every month and tests conducted. The portfolio population is based on the French Insee demographic structure (table RP2009) and includes people between 18 and 62 years-old (see figure 2) for a total of 10^6 individuals (in France, large companies deal with such portfolios).
Numerical results are available in Table 2 to Table 4, and should be read as follows: R is the rejection rate, $E(N)$ is the observed expected sample size and $V(N)$, sample size variance (conditionnaly to rejection). These indicators measure respectively, the power of the test, the test reactivity and finally the variability in test rejection time. In particular, the regular χ^2 test (cf. fixed sample size test) is conducted only once at the 12-th month, thus $E(N) = 12$ and $V(N) = 0$ (fixed sample size test should be conducted only once, otherwise the first type error increases).

In Tables 2, 3 and 4, tests are driven under H_0 for two different levels of α (1

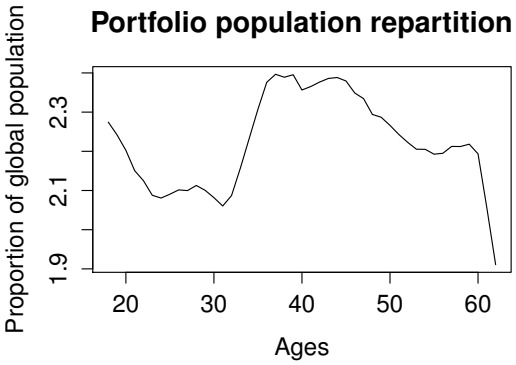

Fig. 2 Population repartition over ages in proportions.

and 5%). We observe that χ^2, χ^2-SPRT are controlled in terms of first-type error probabilities (in each test setups), when the χ^2-Cusum rejects a lot more (cf. Tables 2 and 3). This is due to the fact that quickest detection algorithms are'nt tests as fixed sample or SPRT are. The critical point for this algorithm is the false alarm frequency, which is sufficiently large here (31 months for $\alpha = 5\%$). Nevertheless, when α is small enough, the false alarm frequency increases quickly (11 months for $\alpha = 1\%$). On 12 months setups, Cusum's rejection rate is closer to α, which is clearly a good result.

Table 2 Tests results: $\alpha = 5\%$, $\beta = 0\%$, $\sigma = 0\%$, 1000 simulations, 60 months

	R	$E(N)$	$V(N)$
χ^2	0.05	12.00	0
χ^2-SPRT	0.03	8.81	10.31
χ^2-CUSUM	0.40	30.98	9.80

Table 3 Tests results: $\alpha = 1\%$, $\beta = 0\%$, $\sigma = 0\%$, 1000 simulations, 60 months

	R	$E(N)$	$V(N)$
χ^2	0.02	12.00	0
χ^2-SPRT	0.01	11.00	15.25
χ^2-CUSUM	0.10	10.89	14.11

Table 4 Tests results: $\alpha = 5\%$, $\beta = 0\%$, $\sigma = 0\%$, 1000 simulations, 12 months

	R	$E(N)$	$V(N)$
χ^2	0.04	12.00	0
χ^2-SPRT	0.02	7.15	6.87
χ^2-CUSUM	0.06	7.05	6.16

In case of positive volatility (i.e. H_1 is true), all the tests rejection rates are quickly increasing (see Tables 5, 6 and 7). Cusum power and reactivity is always better than the simple SPRT while the fixed sample size test achieve the best power. However, the interest in this paper is to find tests or algorithms that can monitor death on a continuous basis, avoiding the company to choose when to do the backtest. According to our numerical results, the sequential probability ratio test provides both advantages: a controlled first type error probability and possibility to detect changes at any time.

Table 5 Tests results: $\alpha = 5\%$, $\beta = 0\%$, $\sigma = 10\%$, 1000 simulations, 60 months

	R	$E(N)$	$V(N)$
χ^2	0.92	12.00	0
χ^2-SPRT	1.00	9.65	27.71
χ^2-CUSUM	1.00	8.95	16.11

Table 6 Tests results: $\alpha = 5\%$, $\beta = 0\%$, $\sigma = 20\%$, 1000 simulations, 60 months

	R	$E(N)$	$V(N)$
χ^2	1.00	12.00	0
χ^2-SPRT	1.00	3.69	0.87
χ^2-CUSUM	1.00	3.69	0.86

Table 7 Tests results: $\alpha = 5\%$, $\beta = 0\%$, $\sigma = 10\%$, 1000 simulations, 12 months

	R	$E(N)$	$V(N)$
χ^2	0.93	12.00	0
χ^2-SPRT	0.82	7.79	5.52
χ^2-CUSUM	0.86	7.60	4.99

5 Conclusion

In conclusion of this work, we have presented how statistical modelling, through fixed sample size tests, sequential analysis and change-point detection algorithms can ensure an effective mortality backtesting. Far from being exhaustive, our approach provides fast and simple methods to follow continuously, with controlled first-type error probability and with an acceptable power mortality risk. Indeed, empirical results shows a superior power for fixed sample size tests but they don't provide a suitable practical framework. The sequential probability ratio test is shown to be the most interesting approach for actuaries, keeping a constraint on first type error probability and sample size correlated to the distance between hypothesis and observations. Furthermore, change-point detection algorithms can also be applied to detect shifts in mortality trends. Indeed, we assume that the mortality table is wrong from the start, but both sequential probability ratio test

and cusum algorithm allows for later changes. Finally, we believe that sequential analysis and change-point detection processes can be applied to more complex situations, including disability and multiple other causes. We have to insist that the presented procedures lead to a symmetric appreciation of the tested mortality assumptions. This could and should lead to different consequences depending on whether this lead to an overestimation or underestimation of the predicted risks. Using such techniques enable to get a quantitative appreciation in order to accompany an expert's judgement on the reliability of the mortality assumption.

Acknowledgements This work has been supported by the BNP Paribas Cardif Research Chair "Management de la modélisation en assurance". The authors would like to thank Yahia Salhi for his very hepful advice about detection problems.

A Cox's theorem

Cox's theorem is a powerful tool in sequential analysis. Usually, tests are based on likelihood functions, using observations from the process of interest (here, the number of deaths). Multiple ideas have been developped to take into account vectorial processes, and one is to apply the test directly on statistics (χ^2 or Hotelling-T^2 for example) instead of raw observations. The theorem permits factorization in such a way that the sample probability density function (used for likelihood) reduces to the probability ratio for the statistic (in the following, $l(u_1, ..., u_m, t_1)$ disappears in a probability ratio).

Theorem 1 *Let $x = [x_1, ..., x_n]$ be random variables whose probability density function (p.d.f.) depends on unknown parameters $\theta_1, ..., \theta_p$. The x_i themselves may be vectors. Suppose that:*

- *(i) $t_1, ..., t_n$ are a functionally independent jointly sufficient set of estimators for $\theta_1, ..., \theta_p$,*
- *(ii) the distribution of t_1 involves θ_1 but not θ_2,...,θ_p,*
- *(iii) $u_1, ..., u_m$ are functions of x functionally independent of each other and $t_1, ..., t_p$,*
- *(iv) there exists a set S of transformations of $x = [x_1, ..., x_n]$ into $x^* = [x_1^*, ..., x_n^*]$ such that*
  - *(a) $t_1, u_1, ..., u_m$ are unchanged by all transformations in S,*
  - *(b) the transformation of $t_2, ..., t_p$ into $t_2^*, ..., t_p^*$ is one-to-one,*
  - *(c) if $T_1, ..., T_p$ and $T_2^*, ..., T_p^*$ are two set of values of $t_2, ..., t_p$ each having non-zero probability density under at least one of the distributions of x, then there exists a transformation in S such that if $t_2 = T_2, ..., t_p = T_p$, then $t_2^* = T_2^*, ..., t_p^* = T_p^*$.*

Then the joint p.d.f. of $t_1, u_1, ..., u_m$ factorizes into

$$g(t_1, \theta_1) l(u_1, ..., u_m, t_1), \tag{26}$$

where g is the p.d.f. of t_1 and l doesn't involve θ_1.

B Lorden's theorem

Quickest detection algorithms are stochastic processes, from which we expect some simple properties. Under H_0, the process shouldn't give to frequent false alarm (i.e. $\Delta_0(\tau)$ might be the biggest possible). Under H_1, on the contrary, we wish that the process reacts as quickly as possible, and thus minimize $\Delta_1(\tau)$. The detailed computations of Δ_0 and Δ_1 aren't simple, but Lorden's theorem gives very useful boundaries for practical setups.

Theorem 2 *Let N be a stopping time (or equivalently a sample size) with respect to $y_1, y_2, ...$ such that*

$$P_0(N < \infty) \leq \alpha. \tag{27}$$

For $k = 1, 2, ...$, let N_k be the stopping time obtained by applying N to $y_k, y_{k+1}, ...$ Define the extended stopping time $\tau = \min(k, N_k)$, then:

$$\begin{aligned} \Delta_0(\tau) &\geq \frac{1}{\alpha}, \\ \Delta_1(\tau) &\leq E_1(N). \end{aligned} \tag{28}$$

References

Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Applications*. Prentice-Hall.

El Karoui, N., Loisel, S., Salhi, Y., and Mazza, C. (2013). Fast change detection on proportionnal two-population hazard rates. *Working paper*.

Ghosh, B. K. and Sen, P. K. (1991). *Handbook of Sequential Analysis*. CRC Press.

Gourieroux, C. and Monfort, A. (1996). *Statistique et Modèles économétriques*. Economica.

Hall, W., J., Wijsman, R., A., and Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *The Annals of Mathematical Statistics*, 36:575–614.

Jackson, J., E. and Bradley, R., A. (1961). Sequential chi-2 and t-2 tests. *JST*, pages 1063–1077.

Lai, T., L. (1998). Nearly optimal sequential tests of composite hypotheses. *The Annals of Statistics*, 16:856–886.

Lai, T., L. (2001). Sequential analysis: Some classical problems and new challenges. *Statistica Sinica*, pages 303–408.

Liddell, F. D., K. (1984). Simple exact analysis of the standardised mortality ratio. *Journal of Epidemiology and Community Health*, 38:85–88.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42:1897–1908.

Page, E., S. (1954). Continuous inspection schemes. *Biometrika*, 1-2:100–115.

Planchet, F. and Thérond, P.-E. (2011). *Modélisation statistique des phénomènes de durée : Applications actuarielles*. Economica.

Saporta, G. (2006). *Probabilités, analyses de données et statistiques*. Technip.

Siegmund, D. (1985). *Sequential analysis: Tests and Confidence Intervals*. Springer Verlag.

Tartakovsky, A. G. and Moustakides, G. V. (2010a). Discussion on "quickest detection problems: Fifty years later" by albert n. shiryaev. *Sequential Analysis*, 29:386–393.

Tartakovsky, A. G. and Moustakides, G. V. (2010b). State-of-the-art in bayesian change-point detection. *Sequential Analysis*, 29:125–145.

Wald, A. (1947). *Sequential Analysis*. Dover Phoenix Editions.