Mortality: a statistical approach to detect model misspecification

April 8, 2013

Jean-Charles Croix [a,b], Frédéric Planchet [a,c], Pierre-E. Thérond [a,d]

[a] *Université de Lyon, Université Lyon 1, Institut de Science Financière et d'Assurances, 50 Avenue Tony Garnier, F-69007 Lyon, France.*

[b] *BNP Paribas Cardif*

[c] *Prim'Act*

[d] *Galea & Associés*

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 modelling 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, we'll introduce a specification risk supposing the mortality law true in average but subject to random variations. Finally, the suitability of our method will be assessed within this framework.

Résumé

L'avénement de Solvabilité 2 et de l'évaluation *best estimate* des flux de trésorerie futurs pour le calcul des provisions techniques conduit les assureurs à porter une attention toute particulière aux hypothèses utilisées. C'est particulièrement le cas en matière de lois de mortalité puisque les assureurs sont invités à utiliser des lois *best estimate* en lieu et place des tables prudentes imposées par la réglementation. Dans le cadre de la justification de ces hypothèses, les méthodes de *backtesting*, c'est à dire de validation *ex post* des hypothèses de modélisation, sont mises en avant par le régulateur et sont sujettes à une attention croissante de la part des professionnels comme des universitaires. Dans cet article, nous proposons une démarche (compatible avec les modèles de mortalité paramétriques et non paramétriques) articulée autour de tests statistiques dans le cas d'une loi de mortalité en présence de risque de modèle. Ensuite, un risque de spécification est progressivement ajouté, supposant que le modèle est correct en moyenne mais que la loi résulte de variations aléatoires autour de celui-ci. Enfin, nous étudierons la pertinence de la démarche à l'identification d'une erreur de spécification de la loi de mortalité.

Email addresses: `jean-charles.croix@etu.univ-lyon1.fr` (Jean-Charles Croix), `frederic.planchet@univ-lyon1.fr` (Frédéric Planchet), `pierre@therond.fr` (Pierre-E. Thérond).

1 Introduction

An impressive quantitative step has been made these last 25 years in mortality analysis since the review Pollard (1987). Nowadays, 3 different approaches coexist: Expectation where mortality relies on expert predictions, Extrapolation with statistical regression models (see Booth and Tickle (2008) and Planchet and Therond (2011) for different reviews) and Explanation which considers mortality as the result of identified main causes of death (see Tabeau and al. (1999)) for example).

In practice, the majority of actuarial models for mortality analysis are extrapolative. These regression forms can be classified under 3 families: parametric, semi-parametric and non-parametric. The first set considers that the underlying risk is driven by a finite set of parameters (not being always easy to interpret) and uses graduation (see Haberman and Renshaw (1996)) to fit data. For example, the classical Makeham-Gompertz model considers that death rates are the result of an exponentially age-dependent factor and an accidental rest. In the most recent academic work, models with 3 factors (age, cohort and period) as generalized Lee Carter (see Tabeau et al. (2002) p.16) are being developed. Semi-parametric models such as Cox consider that mortality can be derived from a given hazard function and applied to different population by parametric extensions. The last category is non-parametric, considering that dimension of parametric models is insufficient to fit such erratic and unexpected phenomenon. One can find a review of these specific methods in Tomas (2012).

If extrapolation is the most used method by practitioners and academics, the associated Model risk is too largely neglected. A good practice would challenge underlying hypothesis to historical data as often as possible, that's what Backtesting is about. As extrapolation is a statistical approach, the testing theory should be a good candidate to derive such techniques.

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

In this context, we consider the following simple question: How does an insurer verify that his mortality hypothesis are Best-Estimate ? In a first part, a reminder of mortality analysis and models is provided. These statistical models are adequate foundations to develop and support testing processes that detects if prediction errors are the result of sampling variations or an unidentified trend. According to these models, a first set of tests are reviewed. Among

these, are presented asymptotic tests such as Wald, Score or Likelihood ratio and a few Standardized mortality ratio based tests.

Usual tests are defined for finite sets of observations and not usually repeated. If an insurer repeats such tests on a growing set of data (every month during 3 years for example), the first term error probability converges to one with the number of repetition if no correction is taken on the significance level. That's why a simple method is presented in the third part of this paper. The backtesting processes are designed to periodically compare observation (i.e. experience) to hypothesis. Even if this category of tests allows for future observations to be treated, the total number of trials and the overall significance level should be fixed in advance. A first numerical application presents how all these tests perform when the mortality law is actually best-estimate. The second application is realized under a simulated specification risk.

2 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 global behavior. 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 Therond (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 homogeneous groups an apply statistical modeling. The most classical segmentation is to distinguish people by their age. In this case, the number of deaths D_x at age x among a population of n_x individuals is a binomial random variable. Considering a population between age x_1 and 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 predictions. For a monthly analysis, annual death rates q_x have to be adapted. In this framework, we consider that population under

risk is renewed at every time-steps.Considering that death rates are constant during one year, it's possible to derive monthly mortality rates as follows:

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

${}_mq_x$ being the desired monthly death rate. As a convention in this document, single letters designate vectors over ages (for example, $d = (d_{x_1}, ..., d_{x_p})$ and the subscript x designate a specific age d_x). From a statistical point of view, all mortality models can be considered as parametric models ($\mathcal{Y}$,$\mathcal{P}_Q$) with $\mathcal{Y}$ the set of all possible observations and $\mathcal{P}$ a family of probability distribution on $\mathcal{Y}$ (see Gourieroux and Monfort (1996) for detailed developments). In fact, previous assumptions can be rewritten as the following model:

$$ \mathcal{M}_B = \left(\mathcal{Y} = \mathbb{N}^p, \mathcal{P}_Q = \otimes_{x=x_1}^{x_p} \mathcal{B}\left(n_x, q_x\right) | q \in Q\right), \tag{4} $$

with $Q = [0,1]^p$. If this generic model is well defined, one can find alternatives in the literature (see Planchet and Therond (2011)) for large portfolios, where deaths are approximated using the central limit theorem (see equation A.1 in appendix):

$$ \forall x \in [x_1, x_p],\ \sqrt{n_x}\frac{\hat{q}_x - q_x}{\sqrt{q_x(1-q_x)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{5} $$

Even though this result is asymptotic (i.e. for large n), it's commonly used as the Gaussian law provides ease at use. Again, this new assumption conducts to the following statistical model:

$$ \mathcal{M}_G = \left(\mathcal{Y} = \mathbb{R}^p, \mathcal{P}_Q = \mathcal{N}_p\left(\mu, \Sigma\right) | q \in Q\right), \tag{6} $$

with $\forall x \in [x_1, x_p],\ \mu_x = n_x q_x$ and $(\Sigma)_x = n_x q_x (1 - q_x)$ a diagonal matrix. In the rest of this paper, we'll consider the binomial model $\mathcal{M}_B$ as it's exact for any portfolio size. As S and h are non observable in practice, one can consider multiple mathematical functions to fit data, leading to different sub-models (or mortality models). These functions can be parametric (i.e. defined through different parameters such as age, cohort, period and shape (see Lee Carter 3 factors in Tabeau et al. (2002) for example) or simply assume that mortality should be "smooth" over ages and apply semi-parametric and non-parametric methods (see Tomas (2012) for a detailed presentation).

In this paper, both approaches are treated under a global parametric model. Actually, if one considers $\mathcal{M}_B$ or $\mathcal{M}_G$, the vector of mortality rates q belongs to the set $Q = [0,1]^p$, p being the number of ages considered. Specify a mortality

model can then be interpreted as choosing a subset $\Gamma \in Q$ for mortality rates. In case of parametric sub-models, the subset can be indexed: $\Gamma_\Theta = \{q(\theta)|\theta \in \Theta \in \mathbb{R}^r\}$, with q a differentiable function from Θ (the set of parameters) to Q (corresponding to an explicit hypothesis, see p.570 in Gourieroux and Monfort (1996)). Now that our framework for mortality models is defined, we shall present what our backtesting procedure is.

3 Backtesting a mortality law

Backtesting can be defined as an ex-post model validation method. In statistics, this classical problem is addressed as the decision theory (see Gourieroux and Monfort (1996) or Saporta (2006) for further description). The idea of such method is to reject a modeling hypothesis if observations are statistically unlikely (according to a specific confidence level). Historical observations can be interpreted as a vector d of the possible observations set $\mathcal{Y}$. Supposing that there exists a true probability distribution $\mathcal{P}_0 \in \mathcal{P}_Q$, the backtesting process developed in this paper is designed as a multidimensional composite test with null hypothesis $H_0 = \{\mathcal{P}_0 \in P_\Gamma\}$ (the model is adapted) and its alternative $H_1 = \bar{H}_0$. A backtest is then a mapping ξ from the set of observations $\mathcal{Y}$ to $\{\delta_0, \delta_1\}$ where δ_0 is the decision of keeping H_0 and δ_1 it's opposite (rejecting the model). Alternatively, tests can be defined by their rejection region $W \in \mathcal{Y}$ as they're supra.

In case of $p > 1$ (i.e. multiple ages in the portfolio) and especially for composite tests, optimality theorems can't be easily applied (see Gourieroux and Monfort (1996)) which makes nearly impossible to find a uniformly most powerful test. Our motivation is to present a few and compare their statistical power against a specification risk. Because insurers generally possess large portfolios, asymptotic tests are of primary interest as they're convergent and have asymptotic coverage (definition in Gourieroux and Monfort (1996)).

In practice, it's difficult to test directly $H_0 = \{\mathcal{P}_0 \in \mathcal{P}_\Gamma\}$ in case of non-parametric models because $\mathcal{P}_\Gamma$ is hard to define, our method consists then in testing whether a predefined set of death probabilities q^γ is a Best-estimate of the underlying mortality law q^0 (i.e. $H_0 = \{q^\gamma = q^0\}$) according to observations d. In its implicit form, the previous hypothesis can be rewritten as follows $H_0 = \{g(q) = 0\}$ with $g(q) = q - q^\gamma$ a differentiable function from Q to $\mathbb{R}^p$ (useful formulation in practice). In the following, $\hat{q}$ designates the unconstrained likelihood estimator of q^0 (i.e. $\forall x \in [x_1, x_p]$, $\hat{q}_x = \frac{d_x}{n_x}$ also known as gross mortality rates) and $\mathcal{L}(D, q)$ the model likelihood function.

Since our motivation is operational, we consider the problematic of an insurer motivated to compare his experience against hypothesis as often as possible. In

this case, we consider a mortality portfolio with monthly observations during 3 years (evolution in mortality is neglected during this time). This specific case is treated in a dedicated section.

3.1 Classical significance tests

In this first part, we apply Likelihood based tests, Standardized Mortality Ratio (SMR) based tests and Lindeberg's Central Limit Theorem (see annex equation A.2) alternative tests to our decision problem. All these statistical tests suppose a unique trial, repetition being treated in the next section.

3.1.1 Likelihood based tests (Wald, Score and Likelihood ratio)

In the literature, the most classical asymptotic tests are the Wald, Score and Likelihood ratio. If their formulation differ, they all are convergent, have asymptotic coverage and are asymptotically equivalent (full developments are given in Gourieroux and Monfort (1996)). Wald, Score and Likelihood ratio statistics are all asymptotically chi-squared distributed and defined respectively as ξ^W, ξ^S and ξ^R in this work (notation adopted from Gourieroux and Monfort (1996)).

In particular, the Wald test ensures that the distance between q^γ and q^0 is null (the constraint is verified, $g(\hat{q}) = 0$). The Wald statistic is the following:

$$\begin{aligned} \xi^W &= (\hat{q} - q^\gamma)^t \mathcal{I}(\hat{q}) (\hat{q} - q^\gamma) \\ &= \sum_{x=x_1}^{x_p} \frac{n_x(\hat{q}_x - q_x^\gamma)^2}{\hat{q}_x(1-\hat{q}_x)}. \end{aligned} \tag{7}$$

ξ^W can be seen as a quadratic distance, weighted by the estimated Fisher information. In other words, the more an age has a statistically important information (in sense of Fisher), the more its distance is considered. The chi-squared asymptotic law is obtained considering the projection of a quadratic Gaussian form. It is interesting to remark that a direct application of the classical CLT (with approximate standard deviation) at each age leads to a similar result. In this particular case, the chi-square goodness of fit test statistic with estimated proportions is equal.

The second asymptotic test of this section is the Score test (also known as the Lagrangian multiplier test). The idea of this test is to verify that the likelihood function is at maximum (i.e. the Score is null). This method is particularly

powerful for small deviations. The Score statistic is based on the likelihood function gradient (i.e. the Score) and is computed thereby:

$$\begin{aligned}\xi^S &= \frac{\partial \ln \mathcal{L}(D, q^\gamma)}{\partial q}^t \mathcal{I}^{-1}(q^\gamma) \frac{\partial \ln \mathcal{L}(D, q^\gamma)}{\partial q} \\ &= \sum_{x=x_1}^{x_p} \frac{n_x(\hat{q}_x - q_x^\gamma)^2}{q_x^\gamma(1-q_x^\gamma)}.\end{aligned} \tag{8}$$

In this case, the Score and the Wald statistics are really close. The main difference is that the Score uses the true Fisher information under H_0 while Wald considers an estimator, thus the test coverage should converge faster under H_0. Again, the Chi-square is similar for exact proportions.

Finally, the Likelihood ratio test compares the likelihood function in q^γ and $\hat{q}$. Under H_0, this ratio must be close to 1 as one vector should be a good estimator of the second. The test statistic is the following:

$$\begin{aligned}\xi^R &= 2(\ln \mathcal{L}(\hat{q}) - \ln \mathcal{L}(\hat{q}^\gamma)) \\ &= \sum_{x=x_1}^{x_p} D_x \ln\left(\frac{\hat{q}}{q^\gamma}\right) + (n_x - D_x) \ln\left(\frac{1-\hat{q}}{1-q^\gamma}\right).\end{aligned} \tag{9}$$

All previous statistics are asymptotically chi-square distributed, resulting in the common following rejection region:

$$W = \{\xi > \chi^2_{1-\alpha}(p)\}, \tag{10}$$

p being the number of ages considered in the portfolio, α the level of test significance and $\chi^2_{1-\alpha}(p)$ the chi-square quantile with p-degrees of freedom and $1-\alpha$ level.

3.1.2 Standardized Mortality Ratio significance test

The Standardized Mortality Ratio (SMR - see Liddell (1984) and Rosner (2011)) is defined as the ratio between observed and expected deaths:

$$SMR = \frac{\sum_{i=x_1}^{x_p} D_i}{\sum_{i=x_1}^{x_p} E(D_i)}. \tag{11}$$

If one considers that the number of death can be approximated with a Poisson distribution (assuming a piecewise constant hazard function for example), the

SMR is Poisson distributed (As a sum of independent Poisson variates). Under H_0, the SMR calculation gives:

$$SMR = \frac{\sum_{i=x_1}^{x_p} D_i}{\lambda}, \tag{12}$$

with $\lambda = \sum_{i=x_1}^{x_p} -n_i \ln(1-q_i)$. An exact test (i.e. at finite distance or small portfolios) can be derived (see Liddell (1984)) for the SMR (composite test for the Poisson distribution parameter). The two-sided test consists in computing the following p-values distinguishing cases when the SMR is greater than 1 or not. If the SMR is greater than 1 (i.e. $D > E$, excess of deaths), the test p-value is:

$$p = \min\left[2\left(1 - \sum_{k=0}^{D-1} \frac{e^{-\lambda}\lambda^k}{k!}\right), 1\right], \tag{13}$$

otherwise:

$$p = \min\left[2\sum_{k=0}^{D} \frac{e^{-\lambda}\lambda^k}{k!}, 1\right]. \tag{14}$$

3.1.3 Lindeberg's CLT based tests

In this section, we consider asymptotic approximations using Lindeberg's Central Limit Theorem (see annex). Considering the previously defined SMR, asymptotic results can be obtained for both Binomial and Poisson distributions. Under H_0 and supposing a Poisson distribution the SMR is computed thereby:

$$SMR = \frac{\sum_{i=x_1}^{x_p} D_i}{\lambda}, \tag{15}$$

with $\lambda = \sum_{i=x_1}^{x_p} -n_i \ln(1-q_i)$. With these notations, the following statistic is asymptotically chi-squared distributed (see Rosner (2011) p. 253-254):

$$\xi^{SMRp} = \lambda\,(SMR - 1) = \frac{\left(\sum_{i=x_1}^{x_p} D_i - \lambda\right)^2}{\lambda}. \tag{16}$$

From this idea, one can release the Poisson hypothesis considering the binomial distribution, the SMR is then:

$$SMR = \frac{\sum_{i=x_1}^{x_p} D_i}{\sum_{i=x_1}^{x_p} n_i q_i}. \tag{17}$$

For small death probabilities q, the two different SMR are approximately equals (see Taylor developments). Then, the following statistic is also asymptotically chi-squared distributed:

$$\xi^{SMRb} = \frac{\left(\sum_{i=x_1}^{x_p} D_i - n_i q_i\right)^2}{\sum_{i=x_1}^{x_p} n_i q_i (1-q_i)}. \tag{18}$$

Finally, the rejection regions are defined as follows:

$$W = \{\xi > \chi^2_{1-\alpha}(1)\}. \tag{19}$$

The main difference with the Likelihood based tests seen infra is that the distance is evaluated for the whole portfolio and not age by age. This fact provides a better Gaussian approximation (as all individuals are considered at once) and a faster convergence of the test coverage. This is particularly important for small portfolios.

3.2 Repeated tests on accumulating data

In this section, we consider the problem of an insurer wanting to test each month his mortality law. The context is an insurer wanting to periodically compare his assumptions with experience. In mathematical terms, this method is known as sequential analysis (first developed in Wald (1947), see Sigmund (1985) for a more recent review). Every month, the insurer tests a new null hypothesis ${}_mH_0$ (possibly dependent from previous ones) on the basis of all observations available. In this paper, two situations are investigated. The first is based on independent hypotheses resulting from the consideration the current month information only. This first process achieve low power due to an evident loss of information. On the contrary, the second process considers all available information and gives promising results.

In this paper, the sequential procedure is a repetition of previous tests on varying set of data. The null hypothesis is rejected at the first monthly hypothesis rejection ${}_mH_0$ or after the n-th trial. Let's ${}_kd$ designate deaths occurring during the k-th period. A pure repetition of previous statistical tests every time-step increases the overall first term error probability.

As presented in previous section, single tests produce a controlled first type error probability (may it be asymptotic). For example, the first test on ${}_1d$ gives:

$$P(\xi({}_1d) \in W) = \alpha. \tag{20}$$

If H_0 is rejected, the process stops and there are no more opportunities to have a type 1 error, the test significance is known. On the opposite, if the the first test isn't statistically significant, further tests are proceeded. A process running until the end gives the following overall type 1 error probability (α_G, also called the Family wise error rate):

$$\begin{aligned}\alpha_G &= P(\xi({}_1d) \in W_1) + ... \\ &\quad + P(\xi({}_1d) \notin W_1 \cap \xi({}_1d, {}_2d) \notin W_2 ... \cap \xi({}_1d, ..., {}_nd) \in W_n) \\ &= P(\xi({}_1d) \in W_1 \cup \xi({}_1d, {}_2d) \in W_2 ... \cup \xi({}_1d, ..., {}_nd) \in W_n),\end{aligned} \tag{21}$$

clearly higher than α. The most popular method to control (i.e. to specify an upper bound) this overall first type error probability is the Bonferonni's correction, directly derived from Boole's inequality and assuming equal significance repartition:

$$\alpha \leq \frac{\alpha_G}{n}. \tag{22}$$

This relation also holds for an unequal significance repartition:

$$\sum_{i=1}^{n} \alpha_i \leq \alpha_G, \tag{23}$$

where α_i is the i-th test significance level. If this method is highly conservative, no independence hypothesis between tests is needed. Moreover, this simple method is applicable without knowing all tests outcomes thus adapted to chronological approach (in contrast with all methods based on p-values ordering see Benjamini and Hochberg (1995)). Still, the process needs a predefined maximum number of trials and a specific family wise error rate (FWER).

In the specific case of statistics taking only observations of the current month in account, tests are independent and under equal repartition, the overall first term error probability is a geometric series of ratio $(1 - \alpha)$ and first term α:

$$\begin{aligned}\alpha_G &= P(\xi({}_1d) \in W_1) + ... \\ &\quad + P(\xi({}_1d) \notin W_1) P(\xi({}_2d) \notin W_2) ... P(\xi({}_nd) \in W_n) \\ &= \alpha + ... + \alpha(1 - \alpha)^{n-1} \\ &= 1 - (1 - \alpha)^{12}.\end{aligned} \tag{24}$$

It's also possible to distribute the significance between tests, as a geometric series for example. Naturally, the power of such process is highly reduced in comparison with a unique test on all observations the last month. In return, important deviations can be detected more quickly and countermeasures applied immediately. Finally, this simple process can be applied with all significance tests presented before.

One can notice that Bonferroni's upper bound is really close to the Sidak's method for low significance levels. On the contrary, if the FWER is known in the independent case, the Bonferonni's method only gives an upper bound. The real family wise error rate is smaller but test dependent and hard to compute. Furthermore, the power of a process considering all information at each test is higher than repeated independent tests. In conclusion, even if the FWER isn't clearly specified due to dependent tests, it's controlled for repeated tests.

4 Numerical applications

Two different applications are derived from previous framework. The first consists in repeated tests on independent observations with a known family wise error, the second on all information available each months using Bonferroni's upper bound. Tests are conducted under different levels of specification risk (methodology developed supra) and different portfolio sizes. The testing methodology is presented in figure 1 and repeated 10 000 times.

4.1 Specification risk

Specification risk occurs when the model used to fit data doesn't include the true probability distribution (i.e. H_1 holds). In this case, if q^0 is the real mortality law, q^γ the model and ϵ the error term it comes:

$$q^0 = f(q^\gamma, \epsilon), \tag{25}$$

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

$$\forall x \in [x_1, x_p],\ logit(q_x^0) = logit(q_x^\gamma) + \epsilon_x, \tag{26}$$

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

$$\forall x \in [x_1, x_p],\ q_x^0 = \frac{e^{\epsilon_x} q_x^\gamma}{1 + q_x^\gamma (e^{\epsilon_x} - 1)} - E\left(\frac{e^{\epsilon_x} q_x^\gamma}{1 + q_x^\gamma (e^{\epsilon_x} - 1)} - q^\gamma\right). \tag{27}$$

Finally, an illustration is given of multiple q^0 randomly distributed around q^γ (see figure 2).

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 ? The following table 1 shows correspondence between remaining life expectancy incertitude δ and σ (see annex for detailed computation of δ).

4.2 Data simulation and portfolio structure

Numerical applications are based on a virtual portfolio which structure reflects the French Insee demographic table RP2009 between 18 and 62 years-old for different sizes $N = \sum_x n_x$ (see figure 3).

The initial mortality law is simply generated with a Makeham-Gompertz model adjusted on the TH00-02 mortality table.

4.3 Independent tests

In this first part, different significance tests are applied to reject a miss-specified mortality law at monthly time-step during 3 years. Each test considers only the last month observed deaths thus the tests are independent. The family wise error is fixed through Sidak's hypothesis (regularly distributed significance). A first series of tests is realized without specification risk for different portfolios sizes N, while the second series is applied for different levels of noise.

4.3.1 Different portfolio sizes

First tests (Tables 2, 3 and 4) consist in backtesting a correctly defined law (i.e. setting $q^\gamma = q^0$ or $\sigma = 0$) for different population sizes ($N = \sum_x n_x$) during 36 months. In a first time, a very large portfolio is used ($N = 10^9$) which is never encountered in practice, but still a good numerical example and method to check algorithm implementation. The two others portfolios represent those of a typical large and small companies.

For the very large portfolio (see Table 2), empirical rejection rates are close to the theoretical α_G, it validates algorithm implementation.

For large portfolios (see Table 3), tests using approximation as the Wald are unreliable. Indeed, if the portfolio is important, monthly gross death rates are too small which cause over-rejection rates. When there are no deaths in at least one age, the Likelihood ratio statistic isn't defined which makes it difficult to apply in practice. Concerning the Score test, the coverage is significantly higher than expected, causing too many type 1 errors. Finally, the SMR and the two tests based on the Lindeberg's CLT are the most performing tests.

For small portfolios (see Table 4), all tests loose coverage quality. Likelihood based tests are rejecting null hypothesis almost surely which makes them useless. Lindeberg's CLT based tests results are too important coverages, depending on the desired significance. In case of important α, the coverage can be 2 to 4 times to important while lower values are more violated. For the SMR test, all coverages are smaller than expected. On this difficult case, only CLT based and SMR tests can be use with particular caution. Dependent tests give better results for small portfolios.

In conclusion to this first series of tests, we decide not to consider Wald and Likelihood ratio tests for next series of tests as they are unadapted to our problem.

4.3.2 Different levels of specification risk

The following tests are realized on independent observations for a large portfolio ($N = 10^6$) and different levels for σ.

Globally, empirical rejection rates increase with the level of noise. SMR and CLT based tests are equivalent but their power is really low (84% of rejection for $\sigma = 0.4$, see correspondance table). Concerning the Score test, all rejection rates are more important than all other tests (except for $\sigma = 10\%$) as it starts from a higher level in case of null hypothesis. Even though, the over-important rejection rate makes it less reliable than other tests. Globally, the SMR and CLT based tests demonstrate better results for mortality law backtesting in

case of monthly information.

4.4 Dependent tests on accumulating information

In this section, tests are realized considering each month all information available from the beginning. This situation is typically what an insurer faces in mortality monitoring. The overall significance computation is complex but the Bonferonni's upper bound is used to control it.

4.4.1 Different portfolio sizes

As we've done for the first process, different portfolio sizes are tested with a correctly defined mortality law.

For the very large portfolio, we observe rejection rates far lower than the boundary for asymptotic tests. This result is coherent as the family wise error (FWE) is unknown but bounded. On the contrary, the SMR test crosses the bonferonni's boundary for small α_G values. This last point is natural as the test encounters difficulties for portfolios of large size.

For the large portfolio, all asymptotic rejection rates increase but remain far from the boundary. Concerning the SMR test, rejection rates have decreased, respecting the upper Bonferonni's limit. SMR and CLT tests are relatively closed in coverage for this portfolio.

Equivalently to the first process, the Score test is over-reacting due to a lack of data in case of small portfolios. On the other side, this process gives satisfying results for SMR and CLT based tests in case of small, large and very large portfolios. Considering that information becomes consistent with time, the Score test might even be applicable for large portfolios. Globally, this process shows lower first type errors while assuring the same FWE.

4.4.2 Different levels of specification risk

As previously, the process is tested on a large portfolio with specification risk.

From table 12, we observe that the Score is the most powerful test by far. The rejection is almost certain for $\sigma = 10\%$. All other tests are performing better than in previous process but don't perform as well as the Score do.

The results show that this process is globally more powerful than the previous based on independent observations. For $\sigma = 10\%$, the empirical rejection rate

of SMR and CLT tests are 72% against 31% in the previous method while the Score test allows for almost certain rejections. In particular, we observe that SMR and CLT based tests are equivalent for all tested σ.

5 Conclusion

In conclusion, the mathematical framework presented in first part of this work allows insurer to apply statistical significance tests to mortality law backtesting. Our test benchmark includes asymptotic tests, SMR tests and Lindeberg's CLT based tests but additional tests can be applied to this exercise. Among these, asymptotic tests have shown operational weaknesses for monthly analyses as deaths are rare events. On the contrary, these tests are the most powerful in single testing. In addition of applications consisting in unique tests, we propose two different backtesting processes to monitor model risk in a mortality analysis. These two processes allow the insurer to periodically test if his mortality law is Best-estimate according to a pre-defined significance level. We particularly recommend the process considering all information in combination with the Score test. On the basis of our observations, it's clearly the most powerful process of this paper.

A Central Limit Theorems

This section presents the two most-known versions of the Central Limit Theorem (see Saporta (2006)).

A.1 Classical

Let (X_n) be a sequence of independent random variables with equal expected value μ and standard deviation σ, then:

$$\frac{1}{\sqrt{n}}\left(\frac{X_1 + ... + X_n - n\mu}{\sigma}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{A.1}$$

A.2 Lindeberg

Let (X_n) be a sequence of independent random variables with respective expected value μ_i, standard deviation σ_i and cdf F_{X_i}. If one writes $S_n^2 = \sum_{i=1}^n \sigma_i^2$

and the following (Lindeberg's) condition is verified:

$$\lim_{n\to\infty}\left[\frac{1}{S_n^2}\sum_{i=1}^{n}\int_{|x|>\epsilon S_n} x^2 dF_i(x)\right] = 0, \tag{A.2}$$

it comes:

$$\frac{\sum_i X_i - \mu_i}{S_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{A.3}$$

B Fisher information in Binomial model $\mathcal{M}_B$

Considering model $\mathcal{M}_B$ defined in equation 4, the associated fisher information is computed in this section. From the following log-likelihood function:

$$\ln \mathcal{L}(D,q) = \sum_x D_x \ln(q_x) + (n_x - D_x)\ln(1-q_x), \tag{B.1}$$

the Fisher information matrix can be derived:

$$\begin{aligned}(\mathcal{I}_B(q))_x &= E\left(-\frac{\partial^2 \ln \mathcal{L}(D,q)}{\partial q_x^2}\right)\\ &= \frac{n_x}{q_x(1-q_x)}.\end{aligned} \tag{B.2}$$

C Fisher information in Gaussian model $\mathcal{M}_G$

In the $\mathcal{M}_G$ model, D is multinormal and it's likelihood function is the following:

$$\mathcal{L}(D,q) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(D-\mu)^t\Sigma^{-1}.(D-\mu)\right). \tag{C.1}$$

The Fisher information is computed thereby:

$$
\begin{aligned}
(\mathcal{I}_G(q))_x &= E\left(-\frac{\partial^2 \ln \mathcal{L}(D,q)}{\partial q_x^2}\right) \\
&= \frac{n_x}{q_x(1-q_x)} - \frac{1}{2}\left(\frac{1}{q_x^2} + \frac{1}{(1-q_x)^2}\right) \\
&= (\mathcal{I}_B(q))_x - \frac{1}{2}\left(\frac{1}{q_x^2} + \frac{1}{(1-q_x)^2}\right).
\end{aligned} \tag{C.2}
$$

D Logit noise effect on remaining life expectancy

In order to understand what σ represents in terms of remaining life expectancy, let's consider a 65 years old person. One can compute his remaining life expectancy as follows:

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

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

$$
\delta = \frac{q_{95\%}(e_{65}) - E(e_{65})}{E(e_{65})}. \tag{D.2}
$$

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300.

Booth, H. and Tickle, H. B., L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of actuarial sicence*, 3:3–43.

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

Haberman, S. and Renshaw, A. (1996). Genertrends linear models and actuarial science. *The Statistician*, 45:407–436.

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

Planchet, F. and Therond, P. (2011). *Modelisation statistique des phenomenes de duree: Applications actuarielles*. Ec.

Pollard, J., H. (1987). Projection of age-specific mortality rates. *Population Bulletin of the United Nations*, pages 55–69.

Rosner, B. (2011). *Fundamentals of Biostatistics*. Cengage Learning.

Saporta, G. (2006). *Probabilites analyses de donnees et statistiques*. Technip.

Sigmund, D. (1985). *Sequential analysis: Tests and confidence Intervals*. Springer series in Statistics.

Tabeau, E. and al. (1999). Improving overall mortality forecasts by analysing cause-of-death, period and cohort effects in trends. *European Journal of Population*, 15:153–183.

Tabeau, E., Van Den Berg Jeths, A., and Heathcote, C. (2002). *Forecasting Mortality in Developed Countries: Insight from a Statistical, Demographic and Epidemiological Perspective*. Springer Netherlands.

Tomas, J. (2012). *Quantifying Biometric Life Insurance Risks with Non-Parametric Smoothing Methods*. PhD thesis, Faculty of Economics and Business, University of Amsterdam.

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

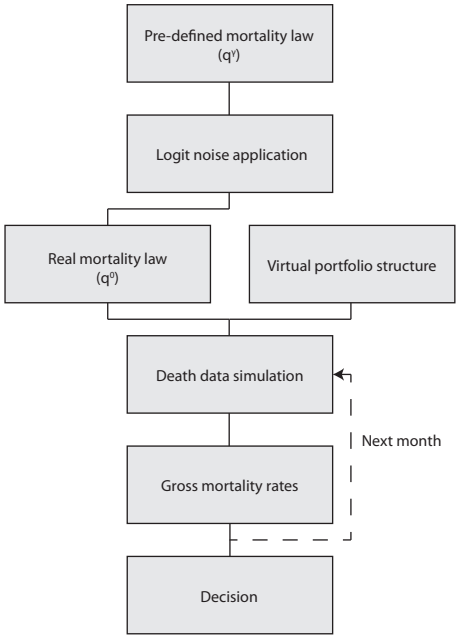

Figure 1. Testing methodology algorithm.

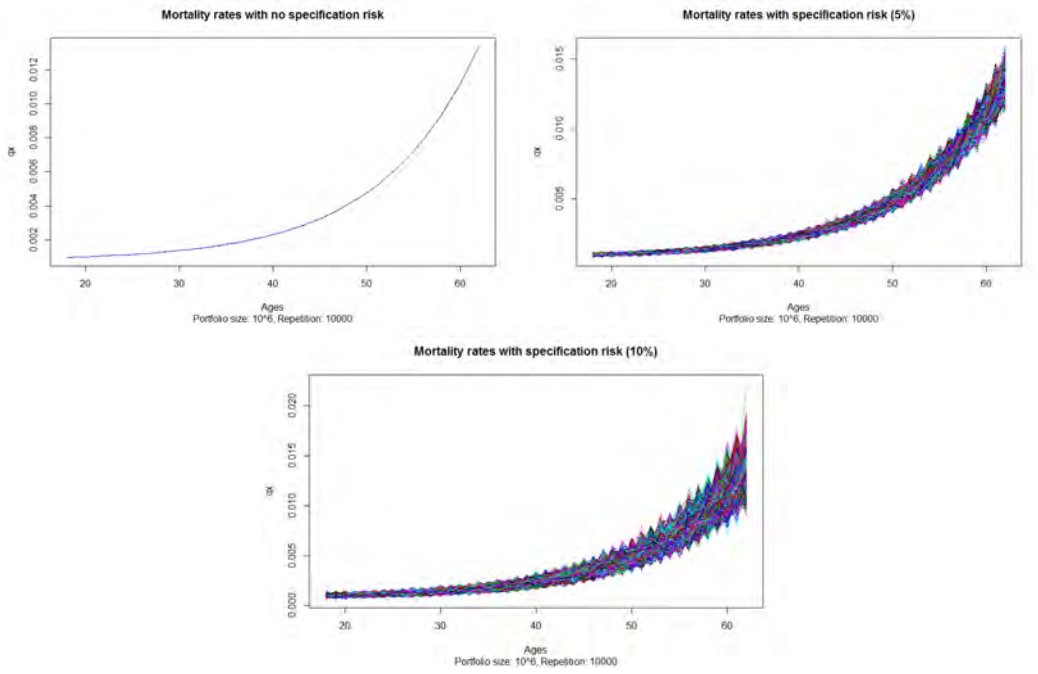

Figure 2. Example of different levels of specification risk $(0, 5\%, 10\%)$.

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

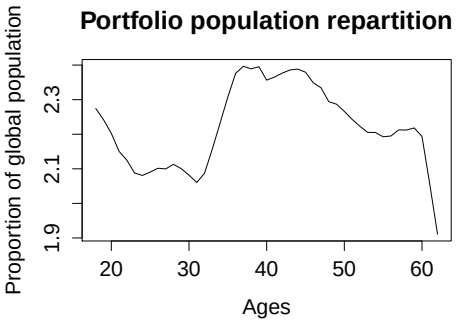

Figure 3. Population repartition over ages in proportions.

Table 2
Empirical rejection rates in case of $H_0 = \{q^\gamma = q^0\}$ and $N = 10^9$.

α_G	10%	5%	1%	0.5%
CLT Binomial	9.87	5.02	1.11	0.50
CLT Poisson	9.85	5.01	1.10	0.49
SMR	10.93	5.69	1.19	0.63
Wald	9.99	4.77	0.92	0.52
Score	9.88	4.70	0.93	0.44
Ratio	9.75	4.71	0.90	0.46

Table 3
Empirical rejection rates in case of $H_0 = \{q^\gamma = q^0\}$ and $N = 10^6$.

α_G	10%	5%	1%	0.5%
CLT Binomial	9.52	4.57	1.03	0.39
CLT Poisson	9.52	4.57	0.85	0.39
SMR	8.66	4.06	0.79	0.34
Wald	100	100	100	100
Score	20.52	12.28	3.97	2.49
Ratio	.	.	.	.

Table 4
Empirical rejection rates in case of $H_0 = \{q^\gamma = q^0\}$ and $N = 10^4$.

α_G	10%	5%	1%	0.5%
CLT Binomial	17.12	17.12	5.79	5.79
CLT Poisson	17.12	17.12	5.79	5.79
SMR	1.72	1.72	0.38	0.08
Wald	100	100	100	100
Score	99.89	99.83	99.30	98.88
Ratio	.	.	.	.

Table 5
Empirical rejection rates in case of H_1, $N = 10^6$ and $\sigma = 10\%$.

α_G	10%	5%	1%	0.5%
CLT Binomial	31.12	18.45	5.02	3.08
CLT Poisson	31.12	18.45	5.02	3.08
SMR	34.02	21.06	6.90	4.05
Score	29.03	18.56	6.65	4.11

Table 6
Empirical rejection rates in case of H_1, $N = 10^6$ and $\sigma = 20\%$.

α_G	10%	5%	1%	0.5%
CLT Binomial	64.50	52.77	30.17	23.52
CLT Poisson	64.50	52.77	30.10	23.52
SMR	66.95	55.45	35.20	27.80
Score	84.63	73.80	48.28	38.75

Table 7
Empirical rejection rates in case of H_1, $N = 10^6$ and $\sigma = 30\%$.

α_G	10%	5%	1%	0.5%
CLT Binomial	78.17	70.70	54.98	48.99
CLT Poisson	78.17	70.70	54.90	48.99
SMR	79.39	72.12	58.60	52.55
Score	99.78	99.35	96.40	93.92

Table 8
Empirical rejection rates in case of H_1, $N = 10^6$ and $\sigma = 40\%$.

α_G	10%	5%	1%	0.5%
CLT Binomial	83.39	78.17	67.11	62.55
CLT Poisson	83.39	78.17	66.94	62.55
SMR	84.03	78.86	69.59	65.20
Score	100	100	100	100

Table 9
Empirical rejection rates in case of H_0 and $N = 10^9$

α_G	10%	5%	1%	0.5%
CLT Binomial	2.44	1.44	0.26	0.18
CLT Poisson	2.44	1.42	0.26	0.18
SMR	6.12	3.88	1.25	0.72
Score	3.16	1.61	0.39	0.19

Table 10
Empirical rejection rates in case of H_0 and $N = 10^6$

α_G	10%	5%	1%	0.5%
CLT Binomial	2.67	1.43	0.42	0.24
CLT Poisson	2.67	1.43	0.41	0.24
SMR	2.63	1.40	0.37	0.23
Score	4.01	2.39	0.66	0.42

Table 11
Empirical rejection rates in case of H_0 and $N = 10^4$

α_G	10%	5%	1%	0.5%
CLT Binomial	3.07	2.00	0.49	0.37
CLT Poisson	3.07	2.00	0.49	0.37
SMR	1.76	0.90	0.24	0.09
Score	32.09	28.56	22.10	19.77

Table 12
Empirical rejection rates in case of H_1, $N = 10^6$ and $\sigma = 10\%$.

α_G	10%	5%	1%	0.5%
CLT Binomial	72.07	68.45	60.16	57.37
CLT Poisson	72.07	68.45	60.13	57.31
SMR	72.58	68.95	60.99	57.95
Score	99.95	99.95	99.79	99.66

Table 13
Empirical rejection rates in case of H_1, $N = 10^6$ and $\sigma = 20\%$.

α_G	10%	5%	1%	0.5%
CLT Binomial	88.30	87.03	83.93	82.72
CLT Poisson	88.30	87.02	83.93	82.71
SMR	88.51	87.28	84.25	82.90
Score	100	100	100	100