A general framework for forecasting numbers of claims

Topic 2 Pricing Risk (Risk Margins)

Author Wright, Thomas, MA, CStat, FIA

Deloitte & Touche LLP
1 Stonecutter Street
London
EC4A 4TR

Tel: +44 (0) 20 7303 6240

thwright@deloitte.co.uk

Abstract

In applications of the collective risk model, significantly more attention is often given to modelling severity than modelling frequency. Sometimes, frequency modelling is neglected to the extent of using a Poisson distribution for the number of claims. The Poisson distribution has variance equal to mean, and there are multiple reasons why this is almost never appropriate when forecasting numbers of non-life insurance claims.

The inappropriateness of the Poisson distribution for forecasting has long been recognised by many, and collective risk algorithms (Panjer (1980), Heckman & Meyers (1983)) have been developed that work just as well with other frequency distributions, in particular the Negative Binomial. However, to calibrate a Negative Binomial model requires two parameters, equivalent to specifying both mean and variance. The author believes that one reason for the prevalence of Poisson models is lack of knowledge about how to objectively quantify the variance as well as the mean. This paper aims to contribute in this area.

The main reasons why the variance should exceed the expected number of claims are identified as parameter estimation error, heterogeneity, contagion, and future exposure uncertainty. While all these factors have long been recognised by some practitioners, this paper provides a framework for their systematic analysis and quantification. A mathematical model is developed in which these concepts are precisely defined, and statistical methods are developed for the quantification of these factors from claim frequency data. The model also shows how these factors interact to produce the overall variance for forecasts.

It is not claimed that the particular form of model presented will be appropriate in all circumstances, but where necessary, modifications will often be possible within the general framework presented here.

Keywords

Bayesian forecasting, claim frequency, collective risk model, contagion, heterogeneity, maximum likelihood estimation, negative binomial distribution, parameter uncertainty, poisson distribution, risk loading.

1 Motivation

1.1 Why is the Poisson assumption 'never' appropriate for forecasting?

1.1.1 Introduction

Claim number probability distributions are used in non-life insurance for forecasting an unknown future number of claims, whether for pricing or reserving purposes. For these purposes, the probability distribution should include all sources of uncertainty, otherwise risk-margins in reserves and/or premiums will be inadequate.

It is useful to distinguish aleatoric and epistemic uncertainty. Aleatoric uncertainty is also known as process uncertainty: epistemic comprises both parameter and model uncertainty. To illustrate these concepts, we consider the example of rolling a normal six-sided die (which, incidentally, is the problem that prompted the invention of probability theory by Pascal and Fermat in 1654).

1.1.2 Process, parameter and model uncertainty: die-rolling example

We aim to predict the number of times a die shows a 'six' when rolled a fixed number of times. If we are sure that the die is perfectly balanced, then the number of 'sixes' follows a Binomial distribution with parameters n and p, where n is the number of throws and p is 1/6. We can use this Binomial distribution to make probabilistic prediction statements, for example: "if we roll the die 4 times, the chance of 4 sixes is 1 in 1,296" (that is, $1/6^4$). The Binomial distribution represents pure aleatoric (or process) uncertainty in this example.

There may also be uncertainty arising from doubt as to whether the die is perfectly balanced. The die may be weighted in such a way as to make a 'six' more or less likely. If it is weighted towards showing a six, then clearly the chance of observing 4 sixes in 4 throws is greater than 1/1,296: if it is weighted against, then the chance of observing 4 sixes is less than 1/1,296. Perhaps less obviously: if we are uncertain as to whether the die is weighted for or against showing a six, then the chance of observing 4 sixes in 4 throws is usually greater than 1/1,296 because of the possibility that the die is weighted in favour.

Suppose, for example, that before observing any throws of the die, we believe there is a 40% chance that it is perfectly balanced (p = 1/6 or 2/12), a 30% chance that in the long-run it will show 1 six in every 12 throws (p = 1/12), and a 30% chance that in the long-run it will show 1 six in every 4 throws (p = ¼ or 3/12).

The chance of obtaining a 'six' in a single throw is now:

0.30 * 1/12 + 0.40 * 2/12 + 0.30 * 3/12 = 1/6, which is the same as for a perfectly balanced die.

However, the chance of 4 sixes in 4 throws is:

$0.30 * (1/12)^4 + 0.40 * (2/12)^4 + 0.30 * (3/12)^4$ = 1/669, which is nearly double the chance of obtaining 4 sixes from a die that is known to be perfectly balanced.

Uncertainty about whether or not the die is perfectly balanced is an example of epistemic uncertainty. This type of epistemic uncertainty is also known as parameter uncertainty: it is uncertainty about the value of the parameter p. The example illustrates that parameter uncertainty generally increases the chances of extreme outcomes.

Another type of epistemic uncertainty is model uncertainty. In the above calculations we assumed that the die has a fixed shape and internal structure so that the chance of throwing a six remains constant. Perhaps the die has a soft centre of non-homogeneous viscous fluid, so as it rests with a six uppermost, its centre of mass shifts down making further sixes more likely. In this case our mathematical model, in which the probability p of a six remains constant, is incorrect. This possibility further increases uncertainty about future outcomes.

1.1.3 Back to insurance: Poisson process

In the die-rolling example, the appropriate probability distribution representing aleatoric uncertainty is the Binomial because the number of sixes is limited by the number of throws. For numbers of non-life insurance claims, a Poisson distribution is usually more appropriate. This is because most non-life policies provide cover for a fixed period of time with no limit on the number of claims. There are exceptions to this, but where there is a contractual limit on the number of claims (as when cover terminates following the first claim, for example) the probability of a claim on each individual policy is usually so low that the Poisson distribution provides a good approximation. (Recall that the Binomial distribution with parameters n and p approaches a Poisson distribution with parameter $\lambda = n.p$ as p tends to zero.)

Some policies do not place a contractual limit on the number of claims but may be subject to practical limits. For example, in auto insurance, a car is usually off the road for repairs following a claim, which creates an upper limit of perhaps one claim every couple of weeks. However, the probability of a claim in any two week period is usually so low that the Poisson distribution again provides a very good approximation.

So for aleatoric uncertainty we use the Poisson distribution. A Poisson distribution is completely specified by its mean, which we denote λ. The Poisson distribution has the property that the variance is equal to the mean. However, in the real-world, the parameter λ is never known precisely, so even if the Poisson model is basically correct, we also have epistemic (parameter) uncertainty. For this reason, the Poisson distribution should not be used for forecasting: we should instead use a distribution with variance greater than mean, such as the Negative Binomial.

1.2 Four types of parameter uncertainty

There are many possible sources of parameter uncertainty in non-life insurance. These are considered in this paper as four main types:

- Estimation uncertainty
- Heterogeneity
- Contagion
- Exposure uncertainty

1.2.1 Estimation uncertainty

Estimation uncertainty is perhaps the purest form of parameter uncertainty. In the real-world it is always present and usually material, so estimation uncertainty alone provides a strong argument for ‘never’ using the Poisson distribution in forecasting claim numbers.

Back to the die-rolling example: suppose we have the opportunity to observe 12 throws of the die before having to forecast the number of sixes that will appear in a further 4 throws. In the 12 throws, we observe 6 sixes. The chance of 6 or more sixes appearing if the die is perfectly balanced can be calculated from the Binomial distribution with $p = 1/6$ as less than 0.8%. We might reasonably conclude from this that p is probably not equal to 1/6. In this case, having observed 6 sixes in 12 throws, we would naturally estimate p as ½. We could now base probabilistic predictions for the next 4 throws on the Binomial distribution with $p = ½$. However, in the absence of any additional ('prior') information, the estimate $p = ½$ is quite unreliable: the standard deviation of this estimate is 0.13 and we can be only 90% confident that the true value of p is between 0.29 and 0.71. (This is a Bayesian confidence interval based on a uniform prior distribution: see Section A.1.1 of the appendix for details.)

Instead of using a single Binomial distribution for forecasting, we should use a mixture of Binomials, allowing for the full range of possible values of the parameter p. The appropriate distribution for forecasting here is known as the 'Beta-Binomial' distribution as it is a mixture of Binomials based on a Beta distribution for p (further details in Section A.1.2 of the appendix). The principles of mixing were illustrated by the numerical example in Section 1.1.2, where we found that parameter uncertainty increases the chance of extreme outcomes.

In the case of a Poisson claim generating process: suppose there were 6 claims last year. In the absence of any other information, we estimate $\lambda=6$ for the annual frequency. The Bayesian posterior distribution for λ is a Gamma distribution, so in forecasting the number of claims next year, we should use a Gamma mixture of Poisson distributions, which is well known to be a Negative Binomial distribution.

Note that throughout this paper, "Negative Binomial distribution" refers to the generalized form in which the parameter r can take any positive real value, not necessarily restricted to integer values. Further details are given in Section A.2.2 of the appendix.

1.2.2 Heterogeneity

Heterogeneity refers to different basic units of exposure, and the common phenomenon that not all are equally risky. In terms of the Poisson model: the parameter λ is usually not the same for all basic risk units. For example, in auto insurance, some drivers are more accident prone than others: in professional indemnity insurance, some lawyers are less scrupulous than others, etc.

Heterogeneity also encompasses situations where the riskiness of a single policy changes with time. For example, a newly qualified driver may become less accident prone during the first year of driving: a professional person may be more likely to make mistakes during busy periods than quiet periods, etc. In these situations, we could consider the basic unit of exposure to be the "policy-month" (or some other suitable time period during which risk might reasonably be approximated as constant): we are then again concerned with the phenomenon that not all exposure units have the same Poisson parameter λ.

"Heterogeneity" can also be stretched to accommodate non-independence of losses arising from a single basic unit of exposure. If one claim tends to lead to another on the same policy, this is much the same situation as heterogeneity over time for a fixed risk unit. (An example of this occurs in household theft insurance where thieves sometimes return for the replacement goods provided by the insurer.) It is also possible that, following a claim, the chance of a further claim from the same exposure unit is temporarily reduced (as when a car is off the road for repairs, for

example). This causes a partial shift towards a Binomial claim generating process, and as we shall see later, can be accommodated as "negative heterogeneity". However, all other causes lead to positive heterogeneity, and when all causes are combined, the overall level of heterogeneity is nearly always positive in non-life insurance.

Note that heterogeneity in its purest form (ie differences between the basic risk units) is only an issue when forecasting if some of the future risk units have not yet been observed. If all risk units have been selected and previously observed (for example, when considering renewal of current policies) heterogeneity does not cause the variance of the number of claims to exceed the mean. This is because, if n_1 and n_2 are Poisson random variables with different parameters λ_1 and λ_2, then their sum $n_1 + n_2$ is also Poisson (with parameter $\lambda_1 + \lambda_2$) so has variance equal to mean. Heterogeneity becomes an issue in forecasting if we expect some new risk units, but don't know whether their Poisson parameters will be λ_1 or λ_2.

1.2.3 Contagion

In this paper, contagion refers to the possibility that the Poisson parameter will temporarily increase (or decrease) for many risk units simultaneously, as a result of a common cause. In many property classes (auto, home-owners, fire, marine etc) a major cause of contagion is the weather. For example, while roads are icy, claim frequency temporarily increases for diverse categories of driver (whether they have high or low Poisson rates in normal conditions). In liability classes, contagion might be caused by a legal ruling that leads to multiple claims.

The difference between heterogeneity and contagion is that heterogeneity refers to differences between exposure units at a particular point in time (or within a particular epoch, such as a calendar year) but contagion refers to differences in risk at different points in time (or between different defined epochs).

Economic conditions are another possible cause of time dependency in many lines of insurance (eg credit insurance, workers compensation). However, these tend to cause gradual trends, or cycles of several years duration, which, to the extent that they are predictable at the start of each year, do not lead to increased parameter uncertainty. Contagion refers to unpredictable time effects: trends are handled separately in the mathematical model developed later in this paper. Trends are unpredictable to the extent that the parameters that describe them have to be estimated, but this is taken into account as parameter estimation uncertainty.

1.2.4 Exposure uncertainty

If λ represents the expected number of claims per unit of exposure, and x represents the expected exposure, then the expected number of claims is λx (assuming uncertainty in these two factors is mutually independent). Clearly some uncertainty in the actual total number of claims is induced by uncertainty in x.

In die-rolling, exposure uncertainty is like not knowing the number of throws n when trying to forecast the number of sixes that will occur. Perhaps we are told only that the die will be thrown repeatedly for one minute. This is analogous to a reinsurance treaty that covers all primary policies in force for one year. In reinsurance treaty pricing, there is often uncertainty in the future level of exposure that will be covered. The extent to which this needs to be explicitly taken into account depends on how the treaty will be priced. If the aim is to produce a premium rate per unit of exposure, then exposure uncertainty has only second order effects.

1.2.5 Summary

At the risk or stating the obvious: each of the four sources of parameter uncertainty described above increases uncertainty when forecasting numbers of claims. If the Poisson distribution is a reasonable model for aleatoric uncertainty in claim numbers, then the presence of any one of these four types of parameter uncertainty will increase the variance so that it exceeds the mean. It is important that all four types of parameter uncertainty are considered otherwise variance may be understated, leading to inadequate risk loads in premiums and/or reserves.

In Section 2 of this paper, we develop a fairly general mathematical model that incorporates Poisson process error and the four types of parameter uncertainty: estimation error, heterogeneity, contagion, and exposure uncertainty. Section 3 outlines statistical methods that can be used to calibrate the model based on past frequency and exposure data. First, in Section 1.3, we give some examples to highlight the importance of taking care with the calibration of the claim number distribution.

1.3 Effect of using wrong claim number distribution

1.3.1 Low frequency example - with estimation error

Consider a catastrophe excess of loss reinsurance treaty covering a specific type of natural catastrophe for one year. The cedent has decided to buy cover with the per event retention set at the level of a 10-year return period. To assess the amount of sideways cover required (the number of reinstatements) the cedent is interested in questions such as: what is the probability of two or more such large events occurring in a single year? The answer to this question clearly depends on the probability distribution assumed for the number of such large events in one year. By definition of events with a 10-year return period, the mean of this distribution is 0.1.

We now suppose that only earthquakes will be covered by the treaty and the cedent carries out calculations using a Poisson distribution in the belief (quite widely supported among seismologists) that this is a reasonable model for earthquake occurrence.

The middle column of Table 1 gives the probability of n events for a Poisson distribution with mean equal to 0.1. The probability that in a single year two or more events occur each producing a loss in excess of the 10-year return value is 0.468% (= 100% - 90.484% - 9.048%).

Table 1 – Probability of n claims – effect of parameter uncertainty

n	No allowance for parameter uncertainty (Poisson)	Allowance for parameter estimation error (Negative Binomial)
0	90.484%	90.555%
1	9.048%	8.913%
2	0.452%	0.509%
3	0.015%	0.022%
4	0.000%	0.001%

However, this calculation ignores parameter uncertainty. The cedent most likely obtained the 10-year return period loss amount by using a catastrophe modelling system: let us suppose this produces a 10-year loss figure for the cedent of $100 million, so this is the per event retention being considered. Because of parameter uncertainty in the cat model, the true annual frequency of quakes causing a gross loss to the cedent in excess of $100 million is not known with certainty to be precisely 0.1. A typical 95% confidence interval for the true frequency would be from 0.038 to 0.192 (corresponding to a return period of between 5.2 and 27 years). When this is taken into account, the variance of the number of events exceeding $100 million increases from 0.1 to 0.1016 (details are given in Section A.2.3 of the Appendix). The chance of two or more losses in excess of $100m increases from 0.468% to 0.532% as shown in the final column of Table 1.

To see what effect this parameter uncertainty should have on technical premiums, we calculate the aggregate loss distribution for a treaty covering the layer $100m excess of $100m each event. We assume that the severity distribution for the part of cat losses in excess of the $100m retention is Pareto with scale $100m and shape parameter 2. This implies that, for each loss event impacting the treaty, the mean amount recoverable is $50m and there is a 25% chance that the full $100m will be recoverable.

The aggregate loss under the treaty has been calculated using both the Poisson and Negative Binomial models for the number of events exceeding $100m gross. Results are shown in columns 2 and 3 of the table below. For both models, the expected number of events is 0.1 and the mean loss to the treaty per event is $50m so the mean aggregate loss is $5m. However the standard deviation of the aggregate loss is slightly higher using the Negative Binomial model. The final two columns relate to the aggregate loss distribution for the same event limits ($100m excess of $100m) but with an aggregate deductible of $100m (in other words, the treaty is a 'back-up' treaty). In this case, the use of a Poisson distribution understates the mean aggregate loss as well as the standard deviation.

Table 2 - Aggregate loss ($m) for treaty covering layer $100m xs $100m per event

	No aggregate deductible		Inner aggregate £100m	
	Poisson	Neg Bin	Poisson	Neg Bin
Mean	$5.00m	$5.00m	$0.11m	$0.12m
Std Dev	$19.65m	$19.76m	$2.72m	$2.95m
Premium	$14.83m	$14.88m	$1.46m	$1.60m

The final row of the table shows a technical premium calculated as mean plus half the standard deviation. In the case of the treaty with no aggregate deductible, it is just the risk load (the standard deviation) that is slightly understated by ignoring estimation error in the frequency parameter, leading to a small, probably immaterial, understatement of technical premium. In the case of the back-up treaty (aggregate deductible of £100m) both the mean and the standard deviation are understated by ignoring parameter estimation error in the frequency, leading to an understatement in the technical premium of approximately 8%.

In the case of events such as hurricanes that have a tendency to occur in clusters, the use of a Poisson distribution would have a more significant impact because it would ignore contagion as well as parameter estimation error. Contagion is considered in the next example.

1.3.2 High frequency example - with contagion

We consider the aggregate loss arising from a home-owners portfolio. The individual loss severity distribution is Log-Normal with mean equal to $1,000 and standard deviation equal to $5,000: this has a skewness coefficient of 140 and implies 1 in 100 losses exceeds $13,000, 1 in 1,000 exceeds $52,000 and 1 in 10,000 exceeds $161,000. We suppose there is a policy limit of $100,000 per claim: this reduces the mean, standard deviation and skewness of the loss distribution to $978.05, $3,632 and 14.1 respectively. The mean claim frequency is estimated as 0.2 per policy year, and we have a portfolio of 1 million policies, so the expected number of claims next year is 200,000.

A Negative Binomial distribution (which has variance greater than mean) is more appropriate than a Poisson distribution (variance equal to mean) for all the reasons discussed in Section 1.2, but contagion will be the dominant cause of super-Poisson uncertainty in this case. The assumption of independent increments required for a Poisson process is clearly violated: claims from a home-owners portfolio do not occur independently because many claims may result from a single cause (often weather related).

Hopefully, nobody would ever use a Poisson distribution in this situation as it is so clearly wrong. The Poisson distribution has standard deviation equal to square root of the mean, so the variation coefficient (defined as the ratio of standard deviation to mean) tends to zero as the mean increases. This is an instance of the 'law of large numbers', which applies here because of the independence assumption implicit in the Poisson assumption. In this example, the mean is 200,000, so the Poisson standard deviation would be 447.2, giving a variation coefficient of only 0.22% for the number of claims. The individual loss distribution (the capped Log-Normal) is so skewed in this example that most of the uncertainty in the aggregate loss would arise from this rather than from variation in the number of claims. (The variation coefficient of the aggregate loss calculated using a Poisson assumption is in fact 0.86%: nearly four times the value attributable to Poisson claim number variation.)

When there are multiple units of exposure as in this example, Poisson parameter uncertainty caused by estimation error or contagion impacts all exposure units in the same direction, so the variation coefficient of the total number of claims does not tend to zero as the size of the portfolio increases. The variation coefficient instead tends to a finite constant determined by the amount of estimation error and contagion. (Here it is assumed that the degree of contagion remains the same as the portfolio grows: if contagion is caused primarily by weather effects, then the degree of contagion will fall if the geographic spread of the portfolio increases.) The general model developed in Section 2 shows that the squared variation coefficient of the number of claims approaches $\rho_e^2 + \rho_c^2 + \rho_e^2.\rho_c^2$ as the portfolio size increases, where ρ_e and ρ_c are the variation coefficients for estimation error and contagion respectively. In this example, we suppose there is no estimation error ($\rho_e = 0$: never true in the real world!) so the variation coefficient for claim numbers approaches ρ_c. The precise definition of ρ_c is elaborated in Section 2: for the time being, the reader is asked to accept that 5% is a realistic value for this example. We therefore use a Negative Binomial distribution with mean equal to 200,000 and standard deviation equal to 10,000 to represent the total number of claims that will occur next year on the portfolio. (In terms of the Negative Binomial parameters used in the appendix: $p = 0.002$, $r = 400.5016$.) Table 3 below gives moments of the aggregate loss distribution calculated using the collective risk model.

Table 3 – Moments of aggregate loss distribution

	Capped Log-Normal	Constant claim severity
Mean	$195,609,000	$195,609,000
Std Deviation	$9,914,450	$9,780,450
Variation Coefficient	5.07%	5.00%
Skewness	0.10002	0.09990
Kurtosis	0.01499	0.01497
Quintessence	1.0032	1.0020
Sixth Moment	0.3257	0.3251

Results in the middle column are based on the capped Log-Normal distribution for individual claim amounts (which has coefficients of variation and skewness of 3.71 and 14.1 respectively). Results in the final column are based on the same Negative Binomial distribution for the number of claims, but a constant claim severity of $978.05: the variation coefficient, skewness and higher moments given in the last column are just those of the Negative Binomial distribution used to model the number of claims.

The striking feature of these results is how close the results based on the Log-Normal severity are to those calculated using a constant severity. This illustrates the fact that *where there is contagion, the shape of the individual loss distribution becomes irrelevant as the expected number of claims increases*. (For a formal proof, see Lundberg (1964).) The same applies where there is parameter estimation or exposure uncertainty. Intuitively, this can be seen as a consequence of the law of large numbers: as the number of claims becomes large, their total approaches their number multiplied by the mean of the severity distribution: variation in the amounts of individual claims cancels out until it is negligible compared to variation in the number of claims. This is not the case in a Poisson model because in that case the coefficient of variation in the number of claims also tends to zero.

The practical consequence is: *in the presence of contagion and/or estimation error and/or exposure uncertainty (ie always), when the expected number of claims is very large, we should not waste our time and effort developing a very refined model of individual loss amounts. We should instead devote most of our effort to ensuring that we have a good model for the number of claims, because the aggregate loss distribution ultimately depends only on this.*

2 General formula for single-epoch forecasts

2.1 Overview

In this section, general formulas for the mean and variance of a claim number forecasting distribution are developed. The formulas take account of all five types of uncertainty identified in Section 1:

- Poisson process uncertainty
- Parameter estimation uncertainty
- Heterogeneity of risk units
- Contagion effects
- Exposure uncertainty

Note that we are concerned in this section only with the first two moments (mean and variance) of the claim number distribution. These are sufficient to calibrate a Negative Binomial distribution for the forecast number of claims. Further comments on the use of a Negative Binomial distribution for this purpose are given in Section 2.3.4.

The number of claims being forecast is denoted N, and its mean and variance are denoted E(N) and Var(N). General formulas for E(N) and Var(N) are derived by repeated application of the following results from the theory of conditional probability:

$$E(X) = E(E(X|Y)) \quad (1)$$

$$Var(X) = E(Var(X|Y)) + Var(E(X|Y)) \quad (2)$$

We start with the Poisson model of aleatoric uncertainty, then apply formulas 1 and 2 once for each of the four types of parameter uncertainty. The final model does not depend on the order in which the four types of parameter uncertainty are taken into account. In the following we introduce them in the order: heterogeneity, exposure uncertainty, contagion, and finally parameter uncertainty.

2.2 Derivation of the general formula

2.2.1 Poisson process uncertainty

As discussed in Section 1.1.3, our basic model of aleatoric uncertainty is the Poisson process. The Poisson parameter may vary between risk units within an epoch (heterogeneity) and may vary from one epoch to the next through contagion effects. So we use λ_{it} to denote the Poisson parameter representing the mean claim frequency for exposure unit i in epoch t.

Initially, we consider a single epoch, and to simplify the notation, we temporarily drop the subscript t. Consider a single risk unit for which the mean frequency λ_i is known. Since we are assuming that the number of claims that will arise, N_i, has a Poisson distribution we have:

$$E(N_i \mid \lambda_i) = Var(N_i \mid \lambda_i) = \lambda_i \quad (3)$$

2.2.2 Heterogeneity

2.2.2.1 Single risk unit drawn at random

Still within a single epoch, imagine that we draw a single risk unit at random from a heterogeneous population of risk units. λ and ρ_h denote the mean and variation coefficient (across all risk units) of the Poisson parameters. λ_i now represents the Poisson parameter of the single randomly selected risk unit: N_i represents the number of claims that will arise from this randomly selected unit. If the population parameters λ, ρ_h are known:

$E(\lambda_i) = \lambda$ and $Var(\lambda_i) = (\rho_h.\lambda)^2$ by definition of λ and ρ_h 4

We now apply formulas 1 and 2 to obtain the mean and variance of N_i:

$E(N_i) = E(E(N_i \mid \lambda_i))$ by equation 1

$= E(\lambda_i)$ by equation 3

$= \lambda$ by equation 4 5

$Var(N_i) = E(Var(N_i \mid \lambda_i)) + Var(E(N_i \mid \lambda_i))$ by equation 2

$= E(\lambda_i) + Var(\lambda_i)$ by equation 3

$= \lambda + (\rho_h.\lambda)^2$ by equation 4

$= \lambda.(1 + \varphi.\lambda)$ where $\varphi = \rho_h^2$ 6

Later we will allow negative values of φ. Clearly if φ is negative, it cannot represent the squared variation coefficient of heterogeneity as defined above. Instead, as discussed in Section 1.2.2, a negative value of φ may arise from non-Poisson aleatoric uncertainty with variance less than mean. If aleatoric uncertainty is better approximated as Binomial, with r being the maximum possible number of claims per risk unit (r is assumed the same for all risk units), then instead of $Var(N_i|\lambda_i) = \lambda_i$ (from equation 3) we have $Var(N_i|\lambda_i) = \lambda_i.(1- \lambda_i/r)$, (because λ_i is now the Binomial mean: $\lambda_i = r.p_i$ say). The above calculation then gives $Var(N_i) = \lambda.(1 + \{\rho_h^2 - (1+ \rho_h^2)/r\}.\lambda)$, which is the same as equation 6 but with $\varphi = \rho_h^2 - (1+ \rho_h^2)/r$. Equation 6 can be viewed as the special case in which r is infinite. For smaller values of r, the combined effects of pure heterogeneity (as represented by ρ_h) and the non-Poisson aleatoric uncertainty may produce a negative value for φ. This is however unusual in non-life insurance.

2.2.2.2 Multiple risk units drawn at random

Now suppose that x units of exposure are selected independently at random from the heterogeneous population, and that each unit generates claims independently. We also assume that the population is very large so the parameters λ and ρ_h of the remaining population hardly change as risk units are withdrawn. N denotes the total number of claims from all x units of exposure. The mean is the sum of the means, and (by independence) the variance is the sum of the variances, so from the above results for a single unit of exposure (equations 5 and 6) we have:

$E(N|x) = x.\lambda$ 7

$Var(N|x) = x.\lambda.(1 + \varphi.\lambda)$ 8

2.2.2.3 Sample drawn partly at random

More generally, suppose a proportion q of the x units of exposure is selected at random from the heterogeneous population, the remainder being a fixed sample with mean Poisson parameter equal to λ. In this case we have:

$$\begin{aligned} \mathrm{Var}(N|x) &= (1-q).x.\lambda + q.x.\lambda.(1 + \varphi.\lambda) \\ &= x.\lambda.(1 + q.\varphi.\lambda) \end{aligned} \qquad 9$$

The first term is for the (1-q).x exposure units that are fixed with average Poisson rate equal to λ. Even though there may be heterogeneity within this group, the total number of claims is the sum of the Poisson numbers generated by each unit of the group, so is also Poisson with parameter equal to the sum of the individual Poisson parameters, that is (1-q).x times the mean Poisson parameter. Heterogeneity within this fixed part of the portfolio affects the reliability with which the mean Poisson rate λ can be estimated from past data for the same portfolio, but this is a separate issue taken into account later (parameter estimation error, Section 2.2.5).

Equation 9 approximates the usual situation in experience-rating applications of the collective risk model: we are pricing for a future period having observed claim numbers in past periods. The past data are used to estimate λ. If the future exposure is mostly renewal business, it can usually be expected to have approximately the same underlying parameter λ as in the past (unless there is reason to believe that the lapsing business will not be a representative cross-section of the portfolio). Any new business has not been observed in the past, so there is the additional uncertainty of drawing this from a heterogeneous population.

2.2.3 Exposure uncertainty

We now suppose that there is uncertainty in the level of exposure x: we use m to denote the expected exposure and ρ_x to denote the variation coefficient:

$$E(x) = m \quad \text{and} \quad \mathrm{Var}(x) = (m.\rho_x)^2 \qquad 10$$

Applying equations 1 and 2 again we have:

$$\begin{aligned} E(N) &= E(E(N|x)) && \text{by equation 1} \\ &= E(x.\lambda) && \text{by equation 7} \\ &= m.\lambda && \text{by equation 10} \qquad 11 \\ \mathrm{Var}(N) &= E(\mathrm{Var}(N|x)) + \mathrm{Var}(E(N|x)) && \text{by equation 2} \\ &= E(x.\lambda.(1 + q.\varphi.\lambda)) + \mathrm{Var}(x.\lambda) && \text{by equations 7 and 9} \\ &= m.\lambda.(1 + q.\varphi.\lambda) + (m.\lambda.\rho_x)^2 && \text{by equation 10} \\ &= m.\lambda + (m.\lambda)^2 \{q.\varphi /m + \rho_x^2\} \\ &= m.\lambda + \alpha.(m.\lambda)^2 \qquad 12 \end{aligned}$$

$$\text{where: } \alpha = q.\varphi /m + \rho_x^2 \qquad 13$$

2.2.4 Contagion

We now consider more than one epoch (for example, accident years) so we reintroduce subscript t (for time) to distinguish epochs. In place of λ we now have λ_t, representing the mean (over all units of exposure) of the Poisson rate in epoch t. (The squared variation coefficient for heterogeneity, $\varphi = \rho_h^2$, may also depend on time: this possibility is considered in Section 3.)

We use μ to denote the underlying mean frequency, and ρ_c to denote the variation coefficient caused by contagion effects:

$$E(\lambda_t) = \mu \quad \text{and} \quad Var(\lambda_t) = (\mu.\rho_c)^2 \qquad 14$$

from which we have:

$$\begin{aligned} E(\lambda_t^2) &= E(\lambda_t)^2 + Var(\lambda_t) && \text{by definition of variance} \\ &= \mu^2.(1+ \rho_c^2) && \text{by definition of } \mu \text{ and } \rho_c \qquad 15 \end{aligned}$$

Contagion effects are assumed to be stochastically independent for different epochs. The possibility of trends is also considered in Section 3: to allow for trends, μ is replaced by μ_t. Contagion refers only to the random component of λ_t that occurs in addition to any trend changes. Nevertheless, the assumption of stochastic independence of contagion effects across epochs is a strong assumption and should be critically considered in any multi-epoch application of the model developed here.

Introducing subscript t, and making the conditioning on λ_t explicit, equations 11 and 12 become:

$$E(N_t|\, \lambda_t) = m.\lambda_t \qquad 16$$

$$Var(N_t|\, \lambda_t) = m.\lambda_t + \alpha.(m.\lambda_t)^2 \qquad 17$$

Applying equations 1 and 2 again:

$$\begin{aligned} E(N_t) &= E(E(N_t|\, \lambda_t)) && \text{by equation 1} \\ &= E(m.\lambda_t) && \text{by equation 16} \\ &= m.\mu && \text{by equation 14} \qquad 18 \\ Var(N_t) &= E(Var(N_t|\, \lambda_t)) + Var(E(N_t|\, \lambda_t)) && \text{by equation 2} \\ &= E(m.\lambda_t + \alpha.(m.\lambda_t)^2) + Var(m.\lambda_t) && \text{by equations 16 and 17} \\ &= m.\mu + \alpha.m^2.\mu^2.(1+ \rho_c^2) + (m.\mu.\rho_c)^2 && \text{by equation 14} \\ &= m.\mu + (m.\mu)^2 .\{ \alpha.(1+ \rho_c^2) + \rho_c^2\} \\ &= m.\mu + (m.\mu)^2 .\{(1+ \alpha).(1+ \rho_c^2) - 1\} \qquad 19 \end{aligned}$$

2.2.5 Parameter estimation uncertainty

Finally, recognizing that equations 18 and 19 are conditional on the underlying mean frequency μ being known precisely, the uncertainty in μ can be introduced by applying formulas 1 and 2 one more time. Using μ' to represent an unbiased estimate of μ, and ρ_e to denote the variation coefficient of the estimation uncertainty, we have:

$$E(\mu) = \mu' \quad \text{and} \quad Var(\mu) = (\rho_e.\mu')^2 \qquad 20$$

Note that we take a Bayesian perspective here: the estimate μ' is known so is not considered to be a random variable: the true value μ is unknown so is considered to be a random variable.

From equation 20 we have:

$$E(\mu^2) = \mu'^2.(1+\rho_e^2) \qquad 21$$

So finally we have:

$$\begin{aligned} E(N_t) &= E(E(N_t \mid \mu)) && \text{by equation 1} \\ &= E(m.\mu) && \text{by equation 18} \\ &= m.\mu' && \text{by equation 20} \qquad 22 \end{aligned}$$

$$\begin{aligned} Var(N_t) &= E(Var(N_t \mid \mu)) + Var(E(N_t \mid \mu)) && \text{by equation 2} \\ &= E(m.\mu + (m.\mu)^2 .\{(1+ \alpha).(1+ \rho_c^2) - 1\}) + Var(m.\mu) \\ & && \text{by equations 19 and 20} \\ &= m.\mu' + (m.\mu')^2 .(1+\rho_e^2).\{(1+ \alpha).(1+ \rho_c^2) - 1\} + (\rho_e.m.\mu')^2 \\ & && \text{by equation 21} \\ &= m.\mu' + (m.\mu')^2 .\{(1 + \alpha).(1+ \rho_c^2).(1+\rho_e^2) - 1\} \qquad 23 \end{aligned}$$

Equation 23 is the proposed general formula for the variance of the number of claims in a single future epoch, taking account of all the main sources of uncertainty. This can alternatively be expressed in terms of the squared variation coefficient (Vco) which is defined as the variance (equation 23) divided by the square of the mean (equation 22):

$$Vco^2(N_t) = 1/(m.\mu') + \{(1 + \rho_x^2 + \rho_h^2.q/m).(1 + \rho_c^2).(1 + \rho_e^2) - 1\} \qquad 24$$

In equation 24, α has been replaced by its definition (equation 13).

2.3 Comments on the general formula

2.3.1 Comparison with other claim number formulas in actuarial literature

Many texts on risk theory allow for the overall effect of parameter uncertainty but do not distinguish the various causes in a general mathematical model. Here, we consider two well-known works: Heckman and Meyers (1983) and Daykin, Pentikäinen and Pesonen (1994), denoted H&M and DP&P in the following.

H&M discuss contagion, heterogeneity, and parameter estimation error, then use a single parameter c (which they call the contagion parameter) to represent the overall effect through the equation:

$$\text{Var}(N) = E(N).\{1 + c.E(N)\} \quad 25$$

Comparing this to the general formula developed here (equations 22 and 23 above) we see that:

$$c = (1 + \rho_x^2 + \rho_h^2.q/m).(1 + \rho_c^2).(1 + \rho_e^2) - 1 \quad 26$$

so the H&M contagion parameter can be seen as representing the combined effect of the four types of parameter uncertainty considered in the present paper.

DP&P consider what we have called 'contagion', although they use the terms 'short-period oscillations' and 'contamination'. To model this they use a multiplicative 'mixing variable' denoted $\mathbf{q}$ with $E(\mathbf{q}) = 1$, $\text{Var}(\mathbf{q}) = \sigma_q^2$, and arrive at the formula (their equation 2.4.12):

$$\text{Var}(N) = E(N).\{1 + \sigma_q^2.E(N)\} \quad 27$$

In the present paper we model contagion (Section 2.2.4) using: $E(\lambda_t) = \mu$ and $\text{Var}(\lambda_t) = (\mu.\rho_c)^2$. If we define $\mathbf{q}_t = \lambda_t/\mu$, then clearly we have $\lambda_t = \mathbf{q}_t.\mu$ with $E(\mathbf{q}_t) = 1$ and $\text{Var}(\mathbf{q}_t) = \rho_c^2$, from which it is clear that this formulation is equivalent to that used by DP&P, with $\rho_c = \sigma_q$. Comparing equation 27 with the general formula (equation 24) we see that the formula of DP&P is the special case $\rho_x = \rho_h = \rho_e = 0$. DP&P also discuss heterogeneity, and point out that a multiplicative mixing variable can also be used to model this (although they prefer the term 'structure variable' in this case) and they cite Ammeter (1948) and Buhlmann (1970) in this connection. However they make it clear that, for their purposes, σ_q^2 represents contagion only, stating "for most of the applications dealt with in this book the inner variation in the collective is not relevant". They do not explicitly consider the effects of parameter estimation error and exposure uncertainty on claim number forecasts.

By explicitly distinguishing the various causes of parameter uncertainty in a general mathematical model, the present paper aims to ensure that all causes are taken into account so the combined effect is not understated.

2.3.2 Limiting and special cases of the general formula

2.3.2.1 Limit as expected exposure increases

As the expected exposure m increases, with all other quantities constant, the general formula (equation 24) for the variation coefficient of a claim number forecast approaches:

$$\text{Vco}^2(N) = (1 + \rho_x^2).(1 + \rho_c^2).(1 + \rho_e^2) - 1 \quad 28$$

This supports the comment made in the example of Section 1.3.2, where we also assumed $\rho_x = 0$.

2.3.2.2 Limit as parameter uncertainty decreases

If the four component variation coefficients are all much smaller than one (except ρ_h, which need only be small compared to the square root of m/q), then product terms become immaterial and equation 24 becomes approximately:

$$\text{Vco}^2(N) = 1/(m.\mu') + \rho_x^2 + \rho_h^2.q/m + \rho_c^2 + \rho_e^2 \quad 29$$

This is just the sum of five terms, one relating to process uncertainty, and one relating to each of the four types of parameter uncertainty. In general (when the component variation coefficients are not necessarily so small that cross terms can be neglected) the total uncertainty can be considered as the sum of the five terms of equation 29, plus an additional term representing the interactions between the five sources of uncertainty.

2.3.2.3 Case of no contagion, heterogeneity or exposure uncertainty

In the real-world there is always parameter estimation error: that is, we always have $\rho_e > 0$. However, it is possible, though unusual, to have no contagion, no heterogeneity and no exposure uncertainty: that is $\rho_x = \rho_h = \rho_c = 0$. It is interesting to examine the formula in this special case because this gives the smallest possible forecasting variance when aleatoric variation is Poisson. (Smaller forecasting variance is possible if aleatoric uncertainty has lower variance than Poisson because then it is possible for "heterogeneity" to be negative, as discussed in Section 1.2.2.)

The general formula (equation 23) for the variance of the number N of claims that will arise from m exposure units reduces in this special case to:

$$\begin{aligned} \text{Var}(N) &= m.\mu' + (m.\mu')^2 .\rho_e^2 \\ &= m.\mu' + m^2.\text{Var}(\mu') \qquad \text{by definition of } \rho_e \end{aligned} \qquad 30$$

Calibration of the formula in the general case is covered in Section 3. However, in this special case, calibration from experience data is particularly simple, if it can be assumed that $\rho_h = \rho_c = 0$ in the past as well as in the future. Under this assumption, the Poisson parameter is the same for all risk units and in all epochs: $\lambda_{it} = \mu$ for all i and t. So if we observed m_0 risk units and the number of claims was k, then the obvious estimate of μ is $\mu' = k / m_0$. Estimation uncertainty is then $\text{Var}(\mu') = \text{Var}(k) / m_0^2 = \mu / m_0$ (because $\text{Var}(k) = m_0.\mu$ by the Poisson assumption). So (replacing μ with its estimate μ') equation 30 becomes:

$$\begin{aligned} \text{Var}(N) &= m.k / m_0 + m^2.k / m_0^2 \\ &= g.(1+ g).k \qquad \text{where } g = m / m_0 \end{aligned} \qquad 31$$

The first term (g.k) represents Poisson process uncertainty: the second term (g^2.k) represents parameter estimation uncertainty.

To interpret equation 31, we first consider the special case g = 1, that is, both estimation and prediction are based on the same number of risk units. (For example, we have observed one year of experience and aim to predict the number of claims in one future year, in the knowledge that the size of the portfolio will not change.) In this case, equation 31 becomes Var(N) = 2.k. This is easily interpreted by noting that Var(N) represents the expected squared difference between the best estimate k and the actual number N of future claims. Since these are, by assumption, independent Poisson variables, their expected squared difference is the sum of their variances, which is the sum of the two Poisson parameters. Again by assumption, they both have the same Poisson parameter, and the obvious estimate of this is the observed number k, so the sum of the two variances is 2.k.

For g not equal to 1, the obvious estimate of N is g.k, so the Poisson process component of Var(N) becomes g.k. The estimation-error component varies with the square of g because, for a given value of g: $\text{Var}(g.k) = g^2.\text{Var}(k)$.

Note that if g is much greater than one, the Process uncertainty (g.k) may be negligible compared to the estimation uncertainty (g^2.k). Conversely, if g is much less than one, the process uncertainty will dominate and the estimation uncertainty may be negligible: this is illustrated by the example in Section 1.3.1.

The above derivation of equation 31 from equation 30 appeals to intuition on what is a good estimate of μ. More formal arguments can lead to slightly different results, depending on the assumptions made. The appendix (Section A.2.2) gives a Bayesian argument leading to:

$$\text{Var}(N) = g.(1+ g).(k + 1) \quad \text{instead of equation 31} \qquad 32$$

The Bayesian argument given in the appendix also shows that, under certain stated conditions (allowing for process variation and parameter estimation error only), the forecasting distribution of N is precisely Negative Binomial.

2.3.2.4 Case of no parameter estimation uncertainty and no exposure uncertainty

Here we consider variation in the number of claims N caused only by process variation, heterogeneity and contagion. This is needed for the development of the statistical calibration methods described in Section 3.

If there is no exposure uncertainty ($\rho_x = 0$) then the quantity α (defined by equation 13) becomes: $\alpha = q.\varphi /m$. Equation 19 (which does not include parameter estimation uncertainty) then becomes:

$$\text{Var}(N_t) = m.\mu + (m.\mu)^2 .\{(1+ q.\varphi /m).(1+ \rho_c^2) - 1\} \qquad 33$$

This can alternatively be obtained from the general formula (equation 24) by setting $\rho_x = \rho_e = 0$.

2.3.3 Application of the general formula

Equations 22 and 23 give the first two moments of the forecasting distribution for the number of claims N. The first two moments are sufficient to specify a Negative Binomial distribution. An advantage of using the Negative Binomial for forecasting claim numbers is that this enables established compounding methods (those studied by Panjer (1980) and Heckman & Meyers (1983)) to be used in calculating aggregate loss distributions. Using a Negative Binomial is also clearly better than using a Poisson distribution: the Negative Binomial allows two moments (mean and variance) to be correctly specified, the Poisson allows only one moment (mean) to be correctly specified.

In the appendix (Section A.2.2) the Negative Binomial distribution is parameterised using parameters r and p. These parameters are related to the first two moments through:

$$p = E(N) / \text{Var}(N) \quad \text{and} \quad r = E(N).p/(1-p) \qquad 34$$

It was noted in Section 2.3.1 that the general formula for Var(N) (equation 23) can be expressed as $\text{Var}(N) = E(N).\{1 + c.E(N)\}$ where $E(N) = m.\mu'$ and c is given by equation 26. The required Negative Binomial parameters are therefore:

$$p = 1 / \{1 + c.m.\mu'\} \text{ and } r = 1/c \qquad 35$$

Although the Negative Binomial is exact in a few, somewhat artificial, special cases (for example: (a) the case of no contagion, no exposure uncertainty and no heterogeneity discussed in Section A.2.2 of the appendix; and (b) the case of no contagion, no exposure uncertainty, no

parameter estimation uncertainty and Gamma-distributed heterogeneity), the Negative Binomial is in general only an approximation. In general, allowing for all sources of uncertainty produces a mixture of Negative Binomial distributions. (In Section 3.3 one such mixture distribution is developed that allows for process error, heterogeneity and contagion only.) The Negative Binomial provides an adequate approximation in many circumstances but of course it will not always be appropriate. In particular, when the expected number of claims is very large, it is advisable to consider whether some other distribution (eg Beta-Binomial, or a mixture or sum of Negative Binomials) would be better. This is because (as illustrated by the example in Section 1.3.2), as the expected number of claims increases, the aggregate loss distribution approaches the assumed claim number distribution, so this becomes critical.

However, regardless of whether the Negative Binomial or some other distribution is used, the general formula for the variance developed in this paper (equation 23) is equally valid.

3 Calibration of the formula from past data

3.1 Introduction

To apply the general formula for the uncertainty of claim number forecasts (equation 24), it first needs to be calibrated, that is, values need to be assigned to the seven quantities appearing in the formula: the four variation coefficients (ρ_x, ρ_e, ρ_h, ρ_c) and the quantities m, q and μ.

In this section we consider the use of data on past claim frequencies to calibrate the model, on the assumption that the future will be similar to the past. Claim frequency data is of little or no relevance for determining the quantities m, ρ_x and q (representing expected future exposure, uncertainty in exposure, and proportion of exposure that will be new business). We do not consider these three other than to say that they should be based on whatever relevant information is available: q could be estimated by looking at data on past renewal and lapse rates, m and ρ_x by looking past growth rates. For all three, business plans are likely to be more relevant than any past data, and judgement more important than statistical analysis. The statistical methods developed in this section are for estimation of the other four parameters, (μ, ρ_e, ρ_h, ρ_c) from past claim numbers and corresponding exposures.

These parameters will all be subject to estimation error. ρ_e represents the estimation error in μ but will itself be subject to estimation error, as will the estimates of ρ_h and ρ_c. We do not attempt to quantify the estimation error in the quantities ρ_e, ρ_h and ρ_c: this would be to give too much weight to past data. Instead, we develop statistical methods that produce point estimates of these quantities: the methods are based on sound statistical principles (such as maximum likelihood estimation) so the point estimates are believed to be about as reliable as possible. To allow for estimation error, the statistical estimates of ρ_e, ρ_h and ρ_c should be considered and adjusted judgementally if necessary. For example, if the past experience shows little evidence of contagion, the statistical methods developed here will produce a low figure for ρ_c. It could be that large contagion effects are possible, but simply that none have occurred in the recent past. Because of this possibility, an analysis of the data to obtain an objective assessment of the reliability of the point estimate ρ_c would be spurious. No amount of statistical analysis would reliably indicate the potential for contagion effects larger than any that have occurred, so judgement is required. The statistical methods described here provide an objective starting point for the application of judgement.

We restrict attention in this paper to short-tail classes for which actual numbers of claims are known with some certainty for all past exposure periods: the dataset is assumed to consist of the following three items for each data-point:

- past epoch t
- exposure x_{tk}
- corresponding actual number of claims n_{tk}

A triple of values (t, x_{tk}, n_{tk}) is called a data-point or an observation. A dataset will usually contain many data-points, but may contain just one. The subscript k allows for the possibility of more than one data-point for the same epoch. Data-points do not necessarily relate to single 'risk units' (however these might be defined): the number of risk units is indicated by x_{tk}.

If there is more than one observation in any one epoch it is possible to estimate the heterogeneity parameter φ ($=\rho_h^2$). If the data-set also covers more than one past epoch it is possible to estimate the contagion parameter ρ_c. If there is only one observation for each of several past epochs, then it is impossible to distinguish the effects of heterogeneity from the effects of contagion.

There are three main stages in the calibration method developed here:

1. Estimation of λ_t for each past epoch, and φ_t for those epochs with more than one observation. (λ_t denotes the true mean Poisson rate in epoch t including contagion effects: φ_t denotes the heterogeneity parameter for epoch t.)
2. Analysis of changes in frequency across epochs to estimate trend parameters and the contagion parameter ρ_c. We also investigate whether there is evidence that the heterogeneity parameter φ_t varies across epochs, and if not, estimate a constant value, denoted φ.
3. Projection of trends to the required future epoch to determine the expected future claim frequency μ and its estimation uncertainty ρ_e. If Stage 2 shows evidence that φ_t varies with t, we also consider what value of φ is appropriate for the future.

These three stages are described in Sections 3.2, 3.3 and 3.4 respectively.

3.2 Estimation of λ_t and φ_t for each epoch separately

3.2.1 Positive heterogeneity

3.2.1.1 Gamma assumption for heterogeneity

Estimation of λ_t and φ_t is carried out for each epoch separately, so we drop the subscript t initially. From Section 2.2.2 (equations 7 and 8), the number of claims N arising from x randomly selected risk units (with x, λ and φ all known) has:

$$E(N|x,\lambda,\varphi) = x.\lambda \qquad 36$$

$$Var(N|x,\lambda,\varphi) = x.\lambda.(1 + \varphi.\lambda) \qquad 37$$

If we assume that heterogeneity follows approximately a Gamma distribution, then the total Poisson rate for x units of exposure is also Gamma (because the sum of independent identically distributed Gamma variables is also Gamma). The distribution of N is then a Gamma mixture of Poissons, which is Negative Binomial: this well-known result is proved in Section A.2.2 of the

appendix. (Alternatively: from the Gamma assumption for heterogeneity it follows that the number of claims on a single randomly selected risk unit is Negative Binomial, so the number of claims N arising from x units of exposure is also Negative Binomial as the sum of independent identically distributed Negative Binomials.) So, by assuming a Gamma distribution of Poisson rates in the heterogeneous population, we have a Negative Binomial distribution for N:

$$\Pr(N = n \mid x, \lambda, \varphi) = \frac{\Gamma(n+r)}{n!\,\Gamma(r)}(1-p)^n.p^r \quad \text{for some p between 0 and 1, and r > 0.} \qquad 38$$

In general, a Negative Binomial distribution has: p = E(N) / Var(N) and r = E(N) . p / (1-p) (see the appendix) so, from equations 36 and 37 we have:

$$p = 1 / (1 + \varphi.\lambda) \text{ and } r = x / \varphi \qquad 39$$

3.2.1.2 Maximum likelihood estimation of true mean frequency λ for single past epoch

Given data pairs (x_k, n_k), the parameters λ and φ can be estimated by maximum likelihood estimation (MLE): that is, we find the values λ' and φ' that maximise the product over all observations k of the Negative Binomial probabilities (equation 38). (An alternative is to use Bayesian estimation: this is considered in Section 3.2.1.3.) For a single data-pair (x, n), the Negative Binomial log-likelihood L is obtained by taking the natural logarithm of the Negative Binomial probability (equation 38). Writing lg(.) for ln(Γ(.)) this gives:

$$\begin{aligned} L &= \lg(n+r) - \lg(n+1) - \lg(r) + n.\ln(1-p) + r.\ln(p) \\ &= \lg(n+ x / \varphi) - \lg(n+1) - \lg(x / \varphi) + n.\ln(\varphi.\lambda) - (n + x / \varphi).\ln(1 + \varphi.\lambda) \end{aligned} \qquad 40$$

For multiple observations (x_k, n_k) (k = 1, 2,…), the total log-likelihood is the sum over k of the above expression for L. Maximum likelihood estimates of φ and λ are the values that maximise this, and can be found by setting the derivatives $\partial L/\partial\lambda$ and $\partial L/\partial\varphi$ to zero. From equation 40:

$$\begin{aligned} \partial L/\partial\lambda &= n / \lambda - (\varphi.n + x) / (1 + \varphi.\lambda) \\ &= (n - \lambda.x) / \{\lambda.(1 + \varphi.\lambda)\} \end{aligned} \qquad 41$$

If there are multiple observations (x_k, n_k) this must be summed over k before setting to zero. The denominator is the same for all observations, so only the numerator needs to be summed, and setting this to zero gives the maximum likelihood estimate (MLE):

$$\lambda' = (\Sigma_k\, n_k) / (\Sigma_k\, x_k) = n^+ / x^+ \qquad 42$$

where n^+ denotes $\Sigma_k\, n_k$ and x^+ denotes $\Sigma_k\, x_k$ Note that the MLE of λ is just the overall observed mean claim frequency in the epoch concerned. Since n^+ is the number of claims from x^+ units of exposure, equation 37 applies and we have:

$$\mathrm{Var}(n^+) = x^+.\lambda.(1 + \varphi.\lambda) \qquad 43$$

From which:

$$\begin{aligned} \mathrm{Var}(\lambda') &= \lambda.(1 + \varphi.\lambda) / x^+ \\ &= \lambda^2.(\varphi + 1/\lambda) / x^+ \end{aligned} \qquad 44$$

Replacing the unknown quantity λ in this expression by the estimate λ' (equation 42) gives the following approximate expression for the squared variation coefficient of λ':

$$\mathrm{Vco}^2(\lambda') \approx (\varphi + 1/\lambda') / x^+ = \varphi / x^+ + 1 / n^+ \qquad 45$$

3.2.1.3 Bayesian estimation of true mean frequency λ for single past epoch

As an alternative to MLE, we also consider Bayesian estimation. From a Bayesian perspective, MLE is equivalent to using an 'uninformative prior' then estimating the parameter as the mode of the posterior distribution. Here we specify a suitable form of informative prior for λ, then determine the posterior distribution by Bayes' theorem. The best estimate of λ is taken as the mean of the posterior distribution (this is best in the sense of minimising the mean square error): estimation uncertainty is represented by the variation coefficient of the posterior distribution.

Substituting from equations 39 into equation 38, the Negative Binomial probability becomes:

$$\Pr(N = n \mid x, \lambda, \varphi) = \frac{\Gamma(n + x/\varphi).(\varphi.\lambda)^n}{n!\,\Gamma(x/\varphi).(1+\varphi.\lambda)^{n+x/\varphi}} \qquad 46$$

Given multiple data-points (x_k, n_k), the likelihood is the product, over all data-points, of this expression. If π(λ) denotes the prior probability density for the parameter λ, then by Bayes' theorem, the posterior density f(λ) is proportional to the product of this and the factors of the likelihood that depend on λ:

$$f(\lambda) \propto \frac{\pi(\lambda).(\varphi.\lambda)^{n^+}}{(1+\varphi.\lambda)^{n^+ + x^+/\varphi}} \qquad 47$$

For the prior distribution we propose π(λ) proportional to $1/\lambda^\gamma$ for some parameter γ between zero and one. This is an improper prior in the sense that it gives an infinite probability for λ > 0. However if π(λ) is set to zero for λ > Λ, where Λ is suitably large, then π(λ) defines a proper prior distribution for λ in the range of what is considered possible. π(λ) decreases as λ increases, meaning basically that high values of λ are considered less likely than lower values. More precisely: the probability that λ is between λ_0 and $\{\lambda_0^{1-\gamma} + P.\Lambda^{1-\gamma}\}^{1/(1-\gamma)}$ does not depend on λ_0 (it is equal to P for all λ_0). For example, putting γ = ½ and Λ = 10 implies that, a priori, λ has an equal chance (10%) of being between any adjacent two of the following:

zero, 0.1, 0.4, 0.9, 1.6, 2.5, 3.6, 4.9, 6.4, 8.1, 10.

Viewed as a probability distribution for the parameter λ, equation 47 (with $1/\lambda^\gamma$ in place of π(λ)) is a Pearson VI probability density function. In general, the Pearson VI pdf can be expressed:

$$f(\lambda) = \frac{\Gamma(a+b).(\lambda/s)^{a-1}}{s.\Gamma(a).\Gamma(b).(1+\lambda/s)^{a+b}} \qquad 48$$

and the mean and variation coefficient are given by:

$$E(\lambda) = s.a/(b-1) \quad \text{(provided } b > 1) \qquad 49$$

$$Vco^2(\lambda) = \{1 + s/E(\lambda)\}/(b-2) \quad \text{(provided } b > 2) \qquad 50$$

(Note that s is a scale parameter, a and b are shape parameters.)

Comparing equations 47 and 48 we have:

$$s = 1/\varphi, \quad a = n^+ + (1-\gamma), \quad b = x^+/\varphi - (1-\gamma) \qquad 51$$

Therefore (from 49 and 50) the posterior mean and variation coefficient for λ are:

$$E(\lambda) = \{n^+ + (1-\gamma)\}/\{x^+ - (2-\gamma).\varphi\} \qquad 52$$

$$Vco^2(\lambda) = \{1 + s / \lambda'\} / (b-2) \qquad \text{where } \lambda' = E(\lambda)$$

$$= \{\varphi + 1 / \lambda'\} / \{ x^+ - (3-\gamma).\varphi\} \qquad 53$$

Comparing to MLE, we see that the Bayesian mean (equation 52) is higher than the MLE of λ (equation 42). Comparing equations 45 and 53 for the squared variation coefficient: the difference is the presence of the term $-(3-\gamma).\varphi$ in the denominator of the Bayesian formula.

The condition (b>2) for the variance of the posterior distribution to be finite is: $x^+ > (3-\gamma).\varphi$. If this inequality is not satisfied, a finite variance can be calculated by truncating the posterior distribution at the maximum value Λ that is believed to be possible. The comparison of exposure x^+ with heterogeneity φ may at first seem strange because we can change the left side of this inequality just by changing the definition of an 'exposure unit'. For example, if we define an 'exposure unit' to be 10 policy-years instead of a single policy-year, then x^+ will decrease by a factor of 10. However, in this case the mean claim frequency increases by a factor of 10, and the variance in claim frequencies (due to heterogeneity) also increases by the same factor (because, by independence, the variance is the sum of the variances). Therefore, the squared variation coefficient of heterogeneity (φ) decreases by the same factor.

Instead of the improper prior $\pi(\lambda)$ proportional to $1/\lambda^\gamma$, we could use a Pearson VI distribution with scale $s = 1/\varphi$ for the prior: equation 47 still yields a Pearson VI posterior distribution.

3.2.1.4 Estimation of heterogeneity φ from multiple observations (single past epoch)

For the estimation of the heterogeneity parameter we consider only MLE because more general Bayesian methods are relatively intractable.

Differentiating the log-likelihood (equation 40) with respect to the heterogeneity parameter φ, and writing $\Psi(x)$ for the derivative $d\lg(x)/dx$ (where $\lg(x) = \ln(\Gamma(x))$ we have:

$$\varphi^2 . \partial L/\partial\varphi = x. \Psi(x/\varphi) - x. \Psi(n + x/\varphi) + x.\ln(1+ \varphi.\lambda) + \varphi.(n - x. \lambda) / (1 + \varphi.\lambda) \qquad 54$$

It can be shown that this is less than zero for all positive values of φ, so for a single observation (n, x) the MLE of φ is always zero. In the case of multiple observations, there is sometimes a positive value φ' such that the sum over all observations of equation 54 is zero. But sometimes, even with multiple observations, the log-likelihood (equation 40 summed over all observations k) increases indefinitely as φ approaches zero. In this case φ should be set to zero or to a negative value as described in the next sub-section. In any case, the MLE φ' cannot be written in closed form, but when a positive solution exists it can be determined from equation 54 by numerical methods. The function $\Psi(x)$ is known as the digamma function and there are good numerical algorithms for evaluating it.

3.2.2 Negative and zero heterogeneity

3.2.2.1 Negative heterogeneity: Binomial distribution

As discussed in Sections 1.2.2 and 2.2.2.1, "heterogeneity" covers several distinct phenomena, and it is possible that the overall effect is a negative value for φ. Negative heterogeneity can be accommodated using a Binomial distribution for N|x (instead of Negative Binomial).

The Binomial probability function can be expressed:

$$P(N=n) = p^n.(1-p)^{r-n}. \Gamma(r+1) / \{ \Gamma(n+1).\Gamma(r-n+1)\} \qquad 55$$

and the mean and variance are:

$$E(N) = r.p \qquad Var(N) = r.p.(1-p) \qquad 56$$

Equating to equations 36 and 37 gives:

$$p = -\varphi.\lambda \qquad r = -x / \varphi \qquad 57$$

Clearly, from equation 37, we must have $\varphi.\lambda > -1$, which ensures $p < 1$.

Taking natural logs of 55 and substituting from 57, the log-likelihood L is:

$$L = \lg(1 - x / \varphi) - \lg(n + 1) - \lg(1 - n - x / \varphi) + n.\ln(-\varphi.\lambda) - (n + x / \varphi) \,.\, \ln(1 + \varphi.\lambda) \qquad 58$$

The terms containing λ are the same as in the Negative Binomial case (equation 40):

$$L = n.\ln(\lambda) - (n + x / \varphi) \,.\, \ln(1 + \varphi.\lambda) + (\text{terms not involving } \lambda) \qquad 59$$

Equation 59 gives the so-called 'quasi-likelihood' for λ: this can be determined solely from the mean and variance assumptions for n (equations 36 and 37). Because both the Negative Binomial and Binomial log-likelihoods are equivalent to the quasi-likelihood for λ, $\partial L/\partial\lambda$ is the same in both cases (equation 41), so the MLE of λ is the same (equation 42), and the formula for the estimation uncertainty is also the same (equation 45).

To find the MLE of φ, equation 58 has to be differentiated with respect to φ:

$$\varphi^2 . \partial L/\partial\varphi = x.\,\Psi(1 - x / \varphi) - x.\Psi(1-n- x/\varphi) + x.\ln(1+ \varphi.\lambda) + \varphi.(n - x.\,\lambda) / (1 + \varphi.\lambda) \qquad 60$$

This is not the same as in the Negative Binomial case (equation 54). The Binomial MLE of φ is the negative value φ' at which the sum over all observations (n_k, x_k) of this expression is zero. As in the Negative Binomial case, there is not necessarily a solution, and when a solution does exist it cannot be written in closed form but can be found by numerical methods. If a solution exists to both equations 54 and 60, the choice between the positive and negative values of φ can be based on the maximised likelihoods (equations 40 and 58 respectively).

It is also possible to calculate a Bayesian estimate of λ in the Binomial case. The Binomial probability (equation 55) can be regarded as a scaled Beta distribution for λ, and a similar sequence followed as in Section 3.2.1.3 for the Negative Binomial. This is not done here for reasons of space.

3.2.2.2 Zero heterogeneity: Poisson distribution

A zero value for φ corresponds to the Poisson distribution for N (given x and λ):

$$\Pr(N = n \mid x, \lambda) = \exp(-x.\lambda)\frac{(x.\lambda)^n}{n!} \qquad 61$$

From which, the log-likelihood is:

$$L = -x.\lambda + n.\ln(x.\lambda) - \lg(n + 1) \qquad 62$$

Differentiating this with respect to λ and setting to zero, produces the same MLE for λ as in the Negative Binomial and Binomial cases (equation 42).

3.3 Estimation of time effects: trends and contagion

3.3.1 Testing for constant heterogeneity

In general, we may have data covering more than one epoch with data-points (t, x_{tk}, n_{tk}). Equations 36 and 37 continue to apply but in general all quantities may vary with time so a subscript t is required throughout. In the previous section the MLE of λ_t for each epoch t was shown to be $\lambda_t' = n_t^+ / x_t^+$ (where $n_t^+ = \Sigma_k\, n_{tk}$ and $x_t^+ = \Sigma_k\, x_{tk}$). In the case when both φ and λ are the same in all epochs, and all exposure units are randomly selected in each epoch (ie $q = 1$) then all observations can be pooled into a single class and the MLE of λ is $\lambda' = (\Sigma_{tk}\, n_{tk}) / (\Sigma_{tk}\, x_{tk})$.

Equations 54 and 60 can also be summed over all observations (k and t) and solved to find the MLE of φ on the assumption that it is the same in all epochs. This can be done for φ using either the constant estimate λ' or the varying estimates λ_t'. Using the varying estimates λ_t', we can calculate the maximum likelihood (sum of equation 40 or 58 over all observations) based on:

a) a single constant MLE φ'

b) varying MLEs φ_t'

We denote the maximum likelihoods L_a and L_b respectively. The hypothesis that φ is constant can then be tested using Wilks generalized likelihood ratio test: if φ is constant, $2.(L_b - L_a)$ has approximately a chi-squared distribution with T-1 degrees of freedom (where T is the number of different φ_t' estimates in (b)). So if $2.(L_b - L_a)$ could plausibly have come from this chi-squared distribution (if it is less than the 95% point of the distribution, for example) the hypothesis that φ is constant can be accepted, but if $2.(L_b - L_a)$ is implausibly large the hypothesis is rejected.

A likelihood ratio test could be used for the hypothesis that λ_t is constant, but more flexible methods that distinguish different types of variation in λ_t are developed in the next sub-section.

3.3.2 Estimation of trends and contagion

3.3.2.1 Introduction

We hypothesise an underlying mean frequency μ_t that would pertain if there were no contagion effects, and which may vary with time (t) as a result of trends or long-term cycles. Contagion relates to additional short-term effects that cause the actual mean frequency λ_t (across the entire heterogeneous risk population) to differ from μ_t. For contagion we use the model introduced in Section 2.2.4 (equations 14 and 15) but with μ_t in place of μ.

The aim here is to estimate the underlying mean frequency μ_t and the contagion parameter ρ_c from an analysis of the variation in observed frequency across epochs. Variation across epochs in the observed frequencies λ_t' (from equation 42) has three possible causes:

a) Trend changes in the underlying frequency μ_t

b) Contagion effects (causing short-term differences between λ_t and μ_t)

c) Estimation error (differences between the estimates λ_t' and true values λ_t).

We have assessed the magnitude of estimation error (equation 45 or 53): any variation in λ_t' that cannot reasonably be attributed to this will be attributed to causes (a) and (b). Gradual changes are attributed to cause (a), short-term variation (in excess of what can be explained by (c)) is attributed to (b).

3.3.2.2 MLE of trend and contagion parameters

We use the notation $n_t^+ = \Sigma_k\, n_{tk}$ and $x_t^+ = \Sigma_k\, x_{tk}$, that is, n_t^+ denotes the total number of claims observed in epoch t, and x_t^+ denotes the total exposure giving rise to these claims. Since n_t^+ is the number of claims arising from x_t^+ units of exposure, equation 37 applies and we have:

$$\mathrm{Var}(n_t^+) = x_t^+.\lambda_t.(1 + \varphi_t.\lambda_t) \qquad 63$$

Also, assuming that heterogeneity follows approximately a Gamma distribution (as described in Section 3.2.1.1) we have a Negative Binomial distribution for n_t^+ (given x_t^+, φ_t and λ_t), with Negative Binomial parameters p_t and r_t given by:

$$p_t = 1 / (1 + \varphi_t.\lambda_t) \text{ and } r_t = x_t^+ / \varphi_t \qquad 64$$

from which we have:

$$\Pr(n_t^+ \mid x_t^+, \varphi_t, \lambda_t) = \frac{\Gamma(n_t^+ + r_t)}{\Gamma(n_t^+ + 1).\Gamma(r_t)} . \frac{(\varphi_t.\lambda_t)^{n_t^+}}{(1 + \varphi_t.\lambda_t)^{n_t^+ + r_t}} \qquad 65$$

(This is essentially the same as equation 46, with the subscript t to distinguish epochs, and with totals n_t^+ and x_t^+ in place of n and x.)

We aim to estimate the contagion parameter ρ_c defined by equation 14 (Section 2.2.4). We also allow for possible trends by using μ_t in place of μ so equation 14 becomes:

$$E(\lambda_t) = \mu_t \quad \text{and} \quad \mathrm{Var}(\lambda_t) = (\mu_t.\rho_c)^2 \qquad 66$$

A suitable form of model for trends might be:

$$\mu_t = \exp(\beta_0 + \beta_1.t + \beta_2.t^2) \qquad 67$$

The exponentiation ensures that μ_t is greater than zero. Other linear forms could alternatively be used in the exponent.

Given the data pairs (n_t^+, x_t^+) and estimates φ_t obtained as described in Sections 3.2 and 3.3.1, we aim to estimate the contagion parameter ρ_c and the trend parameters $(\beta_0, \beta_1, \beta_2)$. To use MLE, we need to assume a full probability distribution for contagion: we have previously made only second moment assumptions as in equation 14. Since the distribution of n_t^+ given x_t^+, φ_t and λ_t is Negative Binomial (as described by equation 65 above), the distribution of n_t^+ allowing for the uncertainty in λ_t caused by contagion effects is a mixture of Negative Binomials: if $f_t(\lambda_t)$ denotes the probability density function representing contagion effects, we have:

$$\Pr(n_t^+ \mid x_t^+, \varphi_t) = \frac{\Gamma(n_t^+ + r_t)}{\Gamma(n_t^+ + 1).\Gamma(r_t)} . \int_{\lambda_t=0}^{\infty} \frac{(\varphi_t.\lambda_t)^{n_t^+}}{(1 + \varphi_t.\lambda_t)^{n_t^+ + r_t}} . f_t(\lambda_t).d\lambda_t \qquad 68$$

This integral is tractable if we assume that $f_t(\lambda_t)$ is a Pearson-VI distribution with scale parameter equal to $1/\varphi_t$. The Pearson-VI family is sufficiently flexible, having two shape parameters, for the mean and variance to be as required. The Pearson VI probability density can be expressed:

$$f(\lambda) = \frac{\Gamma(a+b).(\lambda / s)^{a-1}}{s.\Gamma(a).\Gamma(b).(1 + \lambda / s)^{a+b}} \qquad 69$$

and the mean and variation coefficient are given by:

$$E(\lambda) = s.a/(b-1) \quad \text{(provided } b > 1) \qquad 70$$

$$Vco^2(\lambda) = (a + b - 1) / \{a.(b-2)\} \quad \text{(provided } b > 2) \qquad 71$$

Note that s is a scale parameter, a and b are shape parameters. If a and b both tend to infinity with a/b, constant the variation coefficient tends to zero while the mean tends to s.a/b.

From equation 66, we require the mean and squared variation coefficient to be μ_t and ρ_c^2. Setting $s = 1/\varphi_t$ then solving for a_t and b_t gives:

$$a_t = \{1 + \varphi_t.\mu_t.(1+ \rho_c^2)\} / \rho_c^2 \text{ and } b_t = \{1 + 2.\rho_c^2 + 1 / (\varphi_t.\mu_t)\} / \rho_c^2 \qquad 72$$

Evaluating the integral in equation 68 gives the following probability distribution for the total number of claims in epoch t (given the total exposure and the heterogeneity parameter):

$$\Pr(n_t^+ \mid x_t^+, \varphi_t) = \frac{\Gamma(n_t^+ + r_t)}{\Gamma(n_t^+ + 1).\Gamma(r_t)} . \frac{\Gamma(a_t + b_t).\Gamma(n_t^+ + a_t).\Gamma(r_t + b_t)}{\Gamma(a_t).\Gamma(b_t).\Gamma(n_t^+ + r_t + a_t + b_t)} \qquad 73$$

This can be used to determine ρ_c and the trend parameters (β_0, β_1, β_2) by MLE. The parameters a_t and b_t depend on ρ_c and (β_0, β_1, β_2) (through equations 72 and 67) but n_t^+ and r_t (equation 64) do not. Taking natural logs, and ignoring terms that do not depend on ρ_c and (β_0, β_1, β_2) gives:

$$L_t = \lg(a_t + b_t) + \lg(n_t^+ + a_t) + \lg(r_t + b_t) - \lg(a_t) - \lg(b_t) - \lg(n_t^+ + r_t + a_t + b_t) \qquad 74$$

Using θ to denote any one of the required parameters (ρ_c β_0, β_1, β_2) we have:

$$\partial L_t/\partial\theta = \{\psi(a_t + b_t) + \psi(n_t^+ + a_t) - \psi(n_t^+ + r_t + a_t + b_t) - \psi(a_t)\}.\partial a_t/\partial\theta + \{\psi(a_t + b_t) + \psi(r_t + b_t) - \psi(n_t^+ + r_t + a_t + b_t) - \psi(b_t)\}.\partial b_t/\partial\theta \qquad 75$$

The partial derivatives $\partial a_t/\partial\theta$ and $\partial b_t/\partial\theta$ are obtained from equations 72 and 67. Equations 74 and 75 have to be summed over all epochs t. The matrix of second derivatives (the Hessian matrix) can be obtained by differentiating a second time. The Newton-Raphson method can then be used to find MLEs of the parameters (ρ_c β_0, β_1, β_2). The approximate variance-covariance matrix of the parameter estimates is obtained as negative the inverse Hessian matrix. Further details are not given here for reasons of space.

3.4 Projection of trends

3.4.1 General case of single-epoch forecast

For single-epoch forecasts, we need to apply equations 22 and 24. This section is concerned with determining appropriate values of μ' and ρ_e^2 for the future epoch concerned. Also, if the likelihood ratio test of Section 3.3.1 shows evidence that φ_t varies with t, it is necessary to consider the appropriate value of φ (denoted ρ_h^2 in equation 24) for the future epoch.

From equation 67 we have:

$$\mu_t = \exp(\beta_0 + \beta_1.t + \beta_2.t^2)$$

If we define:

$$\eta_t = \beta_0 + \beta_1.t + \beta_2.t^2 = (1, t, t^2).\boldsymbol{\beta} \qquad 76$$

(where **β** is the column vector of the β-parameters and the dot denotes matrix multiplication), then we have:

$$\mu_t = \exp(\eta_t) \quad 77$$

Using **β'** to denote maximum likelihood estimates of the β-parameters determined as outlined in the previous sub-section, we can calculate an estimate η_t'of η_t for any future epoch t:

$$\eta_t' = (1, t, t^2).\ \boldsymbol{\beta}' \quad 78$$

At this point we have to consider whether to project any trends forwards (by using the value of time t corresponding to the required future epoch in equation 78). Another option is to assume that the there will be no further trend changes between the latest epoch in the dataset and the future epoch for which forecasts are required. In this case, the value of t used in equation 78 should be that corresponding to the latest epoch in the data. Judgement is obviously required in making this decision. If the estimated trend parameters β_1' and β_2' are small, whether or not the trends are projected may be immaterial. If the values are material, then it is necessary to consider what might have caused the trends in the past: this will help in deciding whether or not the trends are likely to continue in the future. Because we will not necessarily always use the actual chronological time of the future epoch equation 78 it is replaced by:

$$\eta' = (1, t_1, t_2).\boldsymbol{\beta}' \quad 79$$

where t_1 and t_2 denote the values selected to replace t and t^2. Similarly equation 76 becomes:

$$\eta = (1, t_1, t_2).\boldsymbol{\beta} \quad 80$$

Assuming the estimate **β'** is approximately unbiased (on the grounds the it is an MLE and MLEs are asymptotically unbiased) we have $E(\eta') = \eta$. Equation 79 implies the following for the estimation uncertainty in η' arising from estimation uncertainty in **β'**:

$$\mathrm{Var}(\eta') = (1, t_1, t_2).\mathrm{Var}(\boldsymbol{\beta}').(1, t_1, t_2)^T = \sigma_e^2 \text{ say} \quad 81$$

where Var(**β'**) is estimated using negative the inverse Hessian matrix (from equation 74), and T denotes matrix transpose.

By asymptotic normality of MLEs, we approximate the estimates **β'** as multivariate normal. Therefore (from equation 79) the estimate η' is approximately normally distributed and exp(η') is approximately log-normal. Using standard results for the log-normal we have:

$$E(\exp(\eta')) = \exp(\eta + \sigma_e^2 / 2) \quad 82$$

$$\mathrm{Vco}(\exp(\eta')) = \exp(\sigma_e^2) - 1 \quad 83$$

So if we define:

$$\mu' = \exp(\eta' - \sigma_e^2 / 2) \quad 84$$

then using ρ_e^2 to denote the variation coefficient of μ', we have:

$$E(\mu') = \exp(\eta) = \mu \quad \text{by equation 82} \quad 85$$

$$\mathrm{Var}(\mu') = \exp(\sigma_e^2) - 1 \quad \text{by equation 79} \quad 86$$

(The variation coefficient of μ' is the same as the variation coefficient of exp(η') because μ' is just exp(η') multiplied by the known factor $\exp(-\sigma_e^2 / 2)$: we ignore estimation error in σ_e^2).

Equations 84 and 86 give the values μ' and ρ_e^2 that are required for the general single-epoch forecasting formulas (equations 22 and 24).

If the likelihood ratio test of Section 3.3.1 shows evidence that φ_t varies with t, it is necessary to consider the appropriate value of φ (denoted ρ_h^2 in equation 24) for the future epoch. This is probably best left to judgment, based on information on possible causes of heterogeneity, why it might have changed, and future business plans that might affect the mix of risks.

4 Conclusion

For reasons of space, it has not been possible to include in the present paper:

- Numerical examples of the statistical calibration methods described in Section 3.
- Statistical calibration methods for use with long-tail classes in which the ultimate number of claims n_t for past periods is not known for several years. In such cases, model calibration is carried out from a development triangle of claim numbers.
- Generalisation of the single-epoch forecasting formulas developed in Section 2 for multi-epoch forecasting.

The author intends to publish further work in these areas.

5 References

Ammeter (1948): "A generalization of the collective theory of risk in regard to fluctuating basic probabilities", Scandinavian Actuarial Journal 31, 171-198

Bühlmann (1970): "Mathematical methods of risk theory", Springer Verlag

Daykin, Pentikäinen and Pesonen (1994): "Practical risk theory for actuaries", Chapman & Hall

Heckman and Meyers (1983): "The calculation of aggregate loss distributions from claim severity and claim amount distributions", PCAS LXX, 22-61

Lundberg O (1964): "On random processes and their application to sickness and accident statistics"

Panjer (1980): "The aggregate claim distribution and stop loss reinsurance", Transactions of the Society of Actuaries, Vol XXXII, pages 523-545.

Appendix A Bayesian forecasting

This appendix reviews some well-known mathematical results and derives related results that are referred to in Sections 1 and 2.

A.1 Binomial process

A.1.1 Posterior distribution for Binomial parameter p

A Binomial random variable k has probabilities:

$$\Pr(k \mid n, p) = n! / \{k!(n-k)!\}\ p^k .(1-p)^{n-k} \qquad (k = 0, 1, 2, \ldots n)$$

where p is the chance of success in each of n independent trials, and k is the number of successes.

Bayes' theorem gives the posterior probability density for p (having observed k) as:

$f(p|k)$ proportion to: $\pi(p).p^k .(1-p)^{n-k}$ where $\pi(p)$ is the prior distribution

If we have a uniform prior ($\pi(p) = 1$) then $f(p|k)$ is a Beta distribution for p.

More generally, if the prior beliefs can be approximated using a Beta distribution (usually the case as the Beta family is very flexible, having two shape parameters):

$\pi(p)$ proportional to: $p^{\alpha-1} .(1-p)^{\beta-1}$

then the posterior is again a Beta distribuion:

$f(p|k)$ proportional to: $p^{\alpha+k-1} .(1-p)^{\beta+n-(k+1)}$

The mean, mode and variance of a Beta distribution as specified above for $\pi(p)$ are:

$E(p) = \alpha / (\alpha + \beta)$

$\text{mode}(p) = (\alpha - 1) / (\alpha + \beta - 2)$

$\text{Var}(p) = \alpha.\beta / \{(\alpha +\beta)^2.(\alpha +\beta+1)\}$

So the mean and mode of the posterior distribution $f(p|k)$ are:

$E(p|k) = (\alpha+k) / (\alpha+\beta+n)$

$\text{mode}(p|k) = (\alpha+k-1) / (\alpha+\beta+n-2)$

The case of a uniform prior can be considered to be the special case $\alpha = \beta = 1$. In this case, the posterior mean is $(k+1)/(n+2)$ and the posterior mode is k/n (which is the classical maximum likelihood estimate).

A Bayesian confidence interval (sometimes called a "credible interval") is an interval constructed from the posterior distribution. In Section 1.2.1 we used a 90% Bayesian confidence interval for p based on $f(p|k)$ above in the case $\alpha = \beta = 1$, $n = 12$ and $k = 6$.

A.1.2 Forecasting for a Binomial process: Beta-Binomial distribution

When aleatoric uncertainty is Binomial, assuming a Beta prior for the parameter p (eg a uniform prior) implies a Beta posterior distribution for p as described above. Forecasts of future outcomes

from the same process should therefore be based on a mixture of Binomial distributions in which the 'mixing weights' are given by the Beta posterior distribution for p. This is called the 'Beta-Binomial' distribution.

A.2 Poisson process

A.2.1 Posterior distribution for Poisson parameter λ

A Poisson random variable k has probabilities:

$\Pr(k \mid \lambda) = e^{-\lambda}.\ \lambda^k / k! \quad (k = 0, 1, 2, \ldots)$

So by Bayes theorem the posterior distribution for λ (having observed k) is:

$f(\lambda \mid k)$ proportional to: $\pi(\lambda).\ e^{-\lambda}.\ \lambda^k / k!$

If we have an uninformative prior distribution ($\pi(\lambda) = 1$) this is a Gamma distribution for λ, with $E(\lambda) = Var(\lambda) = k+1$.

More generally, if the prior beliefs can be approximated using a Gamma distribution:

$\pi(\lambda)$ proportional to: $e^{-\beta.\lambda}.\ \lambda^{\alpha-1}$

then the posterior is again a Gamma distribution:

$f(\lambda \mid k)$ proportional to: $e^{-(\beta+1).\lambda}.\ \lambda^{\alpha+k-1}$

The mean, mode and variance of a Gamma distribution as specified above for $\pi(\lambda)$ are:

$E(\lambda) = \alpha / \beta$

$\text{mode}(\lambda) = (\alpha-1) / \beta$

$Var(\lambda) = \alpha / \beta^2$

So the mean and mode of the posterior distribution $f(\lambda \mid k)$ are:

$E(\lambda|k) = (\alpha+k) / (\beta+1)$

$\text{mode}(\lambda|k) = (\alpha+k-1) / (\beta+1)$

The case of an uninformative prior can be considered to be the limit as β tends to zero with α fixed at one. The variation coefficient of the prior is then 1, but the mean and standard deviation both tend to infinity. This gives a posterior mean of k+1 and posterior mode k (which is also the classical maximum likelihood estimate of λ).

A.2.2 Forecasting for a Poisson process: Negative Binomial distribution

When aleatoric uncertainty is Poisson, assuming a Gamma prior for the parameter λ implies a Gamma posterior distribution for λ as described above. Forecasts of future outcomes from the same process should therefore be based on a mixture of Poisson distributions in which the 'mixing weights' are given by the Gamma posterior distribution for λ. It is shown below that this produces a Negative Binomial distribution. We assume that k has been observed from a Poisson process with parameter λ, but forecasting is required for a Poisson process with parameter g.λ, that is: $\Pr(n \mid g.\lambda) = e^{-g.\lambda}.\ (g.\lambda)^n / n!$. Initially we assume an uninformative prior (the case $\beta = 0$, $\alpha = 1$), so the posterior distribution is $f(\lambda \mid k) = e^{-\lambda}.\lambda^k / k!$. The mixed distribution appropriate for forecasting is then:

$$\Pr(n) = \int_0^\infty P(n \mid g.\lambda).f(\lambda \mid k).d\lambda$$

$$= \frac{1}{n!.k!}\int_0^\infty \exp(-g.\lambda).(g.\lambda)^n.\lambda^k.\exp(-\lambda).d\lambda$$

$$= \frac{g^n}{n!.k!}\int_0^\infty \exp\{-(g+1).\lambda\}.\lambda^{n+k}d\lambda$$

$$= \frac{g^n}{n!.k!}\frac{\Gamma(n+k+1)}{(g+1)^{n+k+1}}$$

$$= \frac{\Gamma(n+r)}{n!.\Gamma(r)}(1-p)^n.p^r \quad \text{where } p = 1/(g+1) \text{ and } r = k+1$$

This is the Negative Binomial distribution: n can be interpreted as the number of failures before success number r in independent trials each with chance p of success.

The negative Binomial has:

$E(n) = r.(1-p) / p$ $\quad$ $Var(n) = E(n) / p$

So in this case, the forecast number of claims n has:

$E(n) = g.(k+1)$ $\quad$ $Var(n) = g.(1+g).(k+1)$

This should be compared with Equation 30 in Section 2.3.3.3, which was derived from the general model using a non-Bayesian argument.

In the case of an informative Gamma prior, the above derivation is easily modified to produce a Negative Binomial with $p = (1+\beta) / (1+\beta+g)$ and $r = k + \alpha$, and therefore:

$E(n) = g.(k+\alpha) / (1+\beta)$ $\quad$ $Var(n) = g.(1+g+\beta).(k+\alpha) / (1+\beta)^2$

From this we see that the parameter r can take any positive value, not necessarily an integer. When generalized in this way, the Negative Binomial is sometimes called the Polya distribution. The case when r is restricted to positive integers is sometimes called the Pascal distribution.

A Negative Binomial distribution exists for any positive values of mean and variance with variance greater than mean. The parameters p and r are given by:

$p = E(N) / Var(N)$ and $\quad$ $r = E(N).p / (1-p)$

A.2.3 Numerical example of Section 1.3.1

In Section 1.3.1 we have a posterior distribution for λ with a mean of 0.1 and 95% confidence interval (0.038, 0.192). This is corresponds to a posterior Gamma distribution for λ with parameters $\alpha' = 6.25$, $\beta' = 62.5$ (where, in the notation above: $\alpha' = (\alpha+k)$ and $\beta' = (\beta+1)$). The variance of this Gamma distribution is $\alpha' / \beta'^2 = 0.0016$. The variance of the number of claims n is then given by:

$Var(n) = E(Var(n|\lambda)) + Var(E(n|\lambda))$

$= E(\lambda) + Var(\lambda)$ $\quad$ because $n|\lambda$ has a Poisson distribution with parameter λ

$= 0.1 + 0.0016$