



HAL
open science

Modeling frequency distribution above a priority in presence of IBNR

Nicolas Baradel

► **To cite this version:**

Nicolas Baradel. Modeling frequency distribution above a priority in presence of IBNR. Scandinavian Actuarial Journal, 2024, pp.1-16. 10.1080/03461238.2024.2439815 . hal-04903637

HAL Id: hal-04903637

<https://hal.science/hal-04903637v1>

Submitted on 21 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Modeling frequency distribution above a priority in presence of IBNR

Nicolas Baradel*

December 12, 2024

Abstract

In reinsurance, Poisson and Negative binomial distributions are employed for modeling frequency. However, the incomplete data regarding reported incurred claims above a priority level presents challenges in estimation. This paper focuses on frequency estimation using Schnieper's framework [7] for claim numbering. We demonstrate that Schnieper's model is consistent with a Poisson distribution for the total number of claims above a priority at each year of development, providing a robust basis for parameter estimation. Additionally, we explain how to build an alternative assumption based on a Negative binomial distribution, which yields similar results. The study includes a bootstrap procedure to manage uncertainty in parameter estimation and a case study comparing assumptions and evaluating the impact of the bootstrap approach.

1 Introduction

In his profession, a reinsurer has to quote prices for excess of loss covers. Generally, the reinsurer estimates the frequency and severity distributions. For the frequency, the most common choices are the Poisson and Negative binomial distributions. The Poisson distribution can be viewed as the natural distribution in an ideal world: when all claims are independent and occur with a non-random intensity, the distribution is Poisson. However, when there is some uncaptured randomness, the variance is greater than the mean. In the particular case where the intensity of a Poisson distribution follows a Gamma distribution, the overall distribution is known to be a Negative binomial one.

The data the reinsurer receives are often incomplete: only the reported incurred claims above a certain threshold, typically known as the priority. In the context of excess of loss, [7, Schnieper] proposed a model that separates the IBNR (Incurred But Not Reported) into what he termed *true IBNR*: newly reported claims, and *IBNER (Incurred But Not Enough Reported)*: variation in estimated cost over time.

*Inria, CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91200 Palaiseau, nicolas.baradel@polytechnique.edu.

Although [4, Mack] used some of the ideas from the Schnieper’s method, it has not received much attention. Major contributions based on the Schnieper model include [3] and [2]. In the former, the author derives an estimator for the mean square error of the reserves. In the latter, they proposed a non-parametric bootstrap procedure to estimate the distribution of the reserve. Additionally, some valuable insights inspired by Schnieper’s model are found in [5] and [6]. In the former, the author uses Schnieper’s approach to determine the implicit part of the IBNER in the Chain Ladder reserve. In the latter, they adapt the methodology in order to also separate the paid from the incurred claims.

Schnieper also addressed a special case: claim numbering above a priority. The frequency of claims exceeding the priority over time is divided into: claims that newly reach the priority and claims that fall below it. In particular, he proposed assuming a Poisson distribution for claims that reach the priority and a Binomial distribution for claims that drop below it.

In this paper, we focus on frequency estimation in presence of incomplete data, specifically the reported incurred above a priority, using Schnieper’ framework for claim numbering. We show that the total number of claims in his model follows a Poisson distribution at each year of development. Consequently, this framework is consistent with a Poisson distribution for the total number of claims above a priority and provides a consistent framework for parameter estimation. We also propose an alternative assumption based on a Negative binomial distribution, which yields similar results. We show that the total number of claims also follows a Negative binomial distribution for each year of development and we provide an estimation procedure. Additionally, we address claim reserving by providing the distribution of ultimate claim numbers, conditioned on current incurred claims.

The paper is organized as follows. Section 2 presents Schnieper’s general model, with a review of key estimators. Section 3 covers claim numbers above a priority. The first part deals with the Schnieper assumption, from which we derive additional results. Specifically, we obtain the distribution of the total number of claims for both purposes: quotation and reserving. The second part presents an alternative assumption under which we show that the total claim number above a priority follows a Negative binomial distribution. Section 4 describes a bootstrap procedure for each case, addressing uncertainty in parameter estimation. Finally, Section 5 provides a case study comparing assumptions and evaluating the impact of the bootstrap approach and its contribution to the different assumptions.

2 The general model

The Schnieper model, with an aim of excess of loss cover, separates two different behaviors in the IBNR data:

- The occurrence of newly reported claims, which are assumed to arise randomly based on the level of exposure ;
- The progression of previously reported claims, which is determined by the current known amounts.

The Schnieper’s framework requires more summary statistics than the aggregated evolution of the incurred claims, which we introduce below, where $n \geq 1$ denotes the number of years, $1 \leq i \leq n$ represents the occurrence year, and $1 \leq j \leq n$ represents the development year:

- The random variables $(N_{i,j})_{1 \leq i, j \leq n}$ represent the total amount of new excess claims, referring to claims that have not been recorded as excess claims in previous development years ;
- The random variables $(D_{i,j})_{1 \leq i, j \leq n}$ represent the decrease in the total claims amount between development years $j - 1$ and j , concerning claims that were already known in development year $j - 1$.

The $(D_{i,j})_{1 \leq i, j \leq n}$ can be negative in the event of an increase and, by construction, $D_{i,1} = 0$ for all $1 \leq i \leq n$.

Given $(N_{i,j})_{1 \leq i, j \leq n}$ and $(D_{i,j})_{1 \leq i, j \leq n}$, the cumulative incurred data $(C_{i,j})_{1 \leq i, j \leq n}$ can be calculated using the following iterative process:

$$\begin{aligned} C_{i,1} &= N_{i,1}, & 1 \leq i \leq n \\ C_{i,j+1} &= C_{i,j} + N_{i,j+1} - D_{i,j+1}, & 1 \leq i \leq n, 1 \leq j \leq n - 1 \end{aligned} \quad (1)$$

We also introduce non-negative exposures $(E_i)_{1 \leq i \leq n}$ that are assumed to be known and associated with the data mentioned above. Finally, we introduce the following filtration:

$$\mathcal{F}_k := \sigma(N_{i,j}, D_{i,j} \mid i + j \leq k + 1), \quad k \geq 1.$$

The current available information is \mathcal{F}_n . In the context of Schnieper's general model, the following assumption is made:

Assumption 2.1.

H1 The random variables $(N_{i_1,j}, D_{i_1,j})_{1 \leq j \leq n}$ and $(N_{i_2,j}, D_{i_2,j})_{1 \leq j \leq n}$ are independent for $i_1 \neq i_2$.

H2 For $1 \leq j \leq n$, there exists $\lambda_j \geq 0$ and for $1 \leq j \leq n - 1$, there exists $\delta_j \leq 1$ such that

$$\begin{aligned} \mathbb{E}(N_{i,j} \mid \mathcal{F}_{i+j-2}) &= \lambda_j E_i, & 1 \leq i \leq n, \\ \mathbb{E}(D_{i,j+1} \mid \mathcal{F}_{i+j-1}) &= \delta_j C_{i,j}, & 1 \leq i \leq n. \end{aligned}$$

H3 For $1 \leq j \leq n - 1$, there exist $\sigma_j^2 \geq 0$ and $\tau_j^2 \geq 0$ such that

$$\begin{aligned} \text{Var}(N_{i,j} \mid \mathcal{F}_{i+j-2}) &= \sigma_j^2 E_i, & 1 \leq i \leq n, \\ \text{Var}(D_{i,j+1} \mid \mathcal{F}_{i+j-1}) &= \tau_j^2 C_{i,j}, & 1 \leq i \leq n. \end{aligned}$$

The evolution of the $(D_{i,j})_{1 \leq i, j \leq n}$ follows the same process as in Mack's model [4], using incurred claims as the exposure combined with a development factor. However, the new claims generated by the $(N_{i,j})_{1 \leq i, j \leq n}$ represent an additional additive component that depends on the exposure.

From the above assumption, Schnieper introduced the following estimators for the λ 's and the δ 's.

$$\begin{aligned} \widehat{\lambda}_j &:= \frac{\sum_{i=1}^{n-j+1} N_{i,j}}{\sum_{i=1}^{n-j+1} E_i}, \\ \widehat{\delta}_j &:= \frac{\sum_{i=1}^{n-j} D_{i,j+1}}{\sum_{i=1}^{n-j} C_{i,j}}. \end{aligned} \quad (2)$$

They are obviously biasfree estimates of the λ 's and δ 's respectively. Additionally, they are the best linear estimators in $\frac{N_{i,j}}{E_i}$ and $\frac{D_{i,j+1}}{C_{i,j}}$ respectively as a consequence of $H3$.

Schnieper developed a way to estimate the expected value of $C_{n+1,n}$ and the pure premium for the following year, including the total reserve along with its associated mean square error based on estimators of the σ 's and τ 's that he provides. We now present the model for claim number, which can be considered a specific instance of the broader model.

3 The model for claim numbers

Schnieper also dealt with a special case: the number of claims above a priority, which is the focus of this paper. From this point onward, for a given priority level, we define

- The random variables $(N_{i,j})_{1 \leq i,j \leq n}$ represent the number of new excess claims pertaining to accident year i in development year j (were below the priority or not reported in development year $j - 1$) ;
- The random variables $(D_{i,j})_{1 \leq i,j \leq n}$ represent the number of claims that exceeded the priority in development year $j - 1$ but have since decreased in cost to fall below the priority in development year j .

In this context, the $(D_{i,j})$ are now non-negative and remain bounded by $(C_{i,j-1})$. Additionally, all data are integer.

Schnieper proposed that the new claims, denoted as $(N_{i,j})$, follow a Poisson distribution, while the claims decreasing below the priority, represented by $(D_{i,j})$, follow a Binomial distribution. The reasoning is as follows: claims that rise above the priority (either new or previously below the threshold) occur independently at non random intensity, aligning with a Poisson distribution. Meanwhile, each claim above the priority has a certain probability of falling below the threshold each year, occurring independently and leading to a binomial distribution. The next subsection will provide a detailed explanation of these assumptions and additional results as stated in [7], including the finding that $(C_{i,j})$ follows a Poisson distribution at each date.

3.1 The Poisson assumption

The following assumption corresponds to [7, Assumptions $A_1'' - A_2''$]. We add $H1'$ to slightly reinforce $H1$.

Assumption 3.1.

H1' The random variables $(N_{i_1,j}, D_{i_1,j})_{1 \leq j \leq n}$ and $(N_{i_2,j}, D_{i_2,j})_{1 \leq j \leq n}$ are independent for $i_1 \neq i_2$, and the pairs $(N_{i,j}, D_{i,j})$ are independent for all $1 \leq i, j \leq n$.

H2' For $1 \leq j \leq n$, there exists $\lambda_j \geq 0$ and for $1 \leq j \leq n - 1$, there exists $0 \leq \delta_j \leq 1$ such that

$$\begin{aligned} N_{i,j} \mid \mathcal{F}_{i+j-2} &\sim \mathcal{P}(\lambda_j E_i), & 1 \leq i \leq n, \\ D_{i,j+1} \mid \mathcal{F}_{i+j-1} &\sim \mathcal{B}(C_{i,j}, \delta_j), & 1 \leq i \leq n. \end{aligned}$$

Note that $H1'$ implies $H1$, and $H2'$ implies $H2$ and $H3$ from Section 2. $H1'$ adds independance between $N_{i,j}$ and $D_{i,j}$ on the same dates, while $H2'$ specifies the distributions. We assume now that the assumptions $H1'$ and $H2'$ hold. Given these, [7] showed that the $\widehat{\lambda}$'s and $\widehat{\delta}$'s defined in (2) are also the maximum likelihood estimators and are efficient, with their variances given by the inverse of the Fisher information.

From what Schnieper stated in this framework, we shall show additional results. Specifically, under $H1'$ and $H2'$, the $(C_{i,j})_{1 \leq i,j \leq n}$ follow a Poisson distribution. Before stating the result, recall first a classic lemma that will be essential.

Lemma 3.1. *Let N be a random variable with distribution $\mathcal{P}(\lambda)$ with $\lambda > 0$ and D a random variable such that $D | N \sim \mathcal{B}(N, p)$ for $0 < p < 1$. Then*

$$N - D \sim \mathcal{P}(\lambda(1 - p)).$$

Proof. For completeness, we provide the proof of this elementary result . Let $k \in \mathbb{N}$.

$$\begin{aligned} \mathbb{P}(N - D = k) &= \sum_{n \in \mathbb{N}} \mathbb{P}(\{D = n - k\} | \{N = n\}) \mathbb{P}(N = n) \\ &= \sum_{n \geq k} \frac{n!}{k!(n - k)!} p^{n-k} (1 - p)^k \frac{e^{-\lambda} \lambda^n}{n!} \\ &= \frac{e^{-\lambda} [\lambda(1 - p)]^k}{k!} \sum_{n \geq 0} \frac{(p\lambda)^n}{n!} \\ &= \frac{e^{-\lambda(1-p)} [\lambda(1 - p)]^k}{k!}. \end{aligned}$$

□

Proposition 3.2. *For all $1 \leq i, j \leq n$, we have*

$$C_{i,j} \sim \mathcal{P}(\lambda'_j E_i),$$

with

$$\lambda'_j := \sum_{k=1}^j \lambda_k \left(\prod_{\ell=k}^{j-1} (1 - \delta_\ell) \right).$$

Proof. We prove the lemma by induction. Let $1 \leq i \leq n$ be fixed. By construction $C_{i,1} = N_{i,1} \sim \mathcal{P}(\lambda_1 E_i)$. Assume as the induction hypothesis that $C_{i,j}$ follows a Poisson distribution with parameter $\lambda'_j E_i$. Recall the relation in (1):

$$C_{i,j+1} = C_{i,j} + N_{i,j+1} - D_{i,j+1}.$$

Under $H2'$, by Lemma 3.1,

$$C_{i,j} - D_{i,j+1} \sim \mathcal{P}(\lambda'_j E_i (1 - \delta_j)),$$

and finally, under $H1'$:

$$C_{i,j+1} \sim \mathcal{P}(\lambda_{j+1} E_i + \lambda'_j (1 - \delta_j) E_i) = \mathcal{P}(\lambda'_{j+1} E_i).$$

□

The above result provides the distribution of $C_{n+1,n}$ given the corresponding exposure. In practice, we are also interested in the conditional distribution $C_{i,j} \mid \mathcal{F}_n$ for $i+j > n+1$. Before presenting this result, we introduce an essential lemma.

Lemma 3.3. *Let $N \in \mathbb{N}^*$. Let $D_1 \sim \mathcal{B}(N, p_1)$, and $D_{n+1} \mid \{\sum_{i=1}^n D_i = d\} \sim \mathcal{B}(N-d, p_{n+1})$ for $n \geq 1$. Then*

$$N - \sum_{i=1}^n D_i \sim \mathcal{B}\left(N, \prod_{i=1}^n (1 - p_i)\right).$$

Proof. We prove the lemma by induction. It is clear that $N - D_1 \sim \mathcal{B}(N, 1 - p_1)$. Let $n \geq 1$, and assume by induction that

$$N - \sum_{i=1}^n D_i \sim \mathcal{B}\left(N, \prod_{i=1}^n (1 - p_i)\right).$$

Let $k \in \{0, \dots, N\}$. For ease of notation, we introduce $1 - \pi_n := \prod_{i=1}^n (1 - p_i)$. It follows that

$$\begin{aligned} \mathbb{P}\left[N - \sum_{i=1}^{n+1} D_i = k\right] &= \sum_{d=0}^N \mathbb{P}\left[N - \sum_{i=1}^{n+1} D_i = k \mid \sum_{i=1}^n D_i = d\right] \mathbb{P}\left(\sum_{i=1}^n D_i = d\right) \\ &= \sum_{d=0}^N \mathbb{P}\left[D_{n+1} = N - d - k \mid \sum_{i=1}^n D_i = d\right] \mathbb{P}\left(N - \sum_{i=1}^n D_i = N - d\right) \\ &= \sum_{d=0}^{N-k} \frac{(N-d)!}{k!(N-d-k)!} p_{n+1}^{N-d-k} (1-p_{n+1})^k \frac{N!}{d!(N-d)!} \pi_n^d (1-\pi_n)^{N-d} \\ &= \frac{N!}{k!(N-k)!} (1-\pi_{n+1})^k \sum_{d=0}^{N-k} \frac{(N-k)!}{(N-d-k)! d!} p_{n+1}^{N-d-k} \pi_n^d (1-\pi_n)^{N-d-k} \\ &= \frac{N!}{k!(N-k)!} (1-\pi_{n+1})^k (\pi_n + p_{n+1}(1-\pi_n))^{N-k} \\ &= \frac{N!}{k!(N-k)!} (1-\pi_{n+1})^k \pi_{n+1}^{N-k}. \end{aligned}$$

□

Proposition 3.4. *For all $1 \leq i, j \leq n$ such that $i+j > n+1$, we have*

$$C_{i,j} \mid \mathcal{F}_n \sim \mathcal{B}(C_{i,n-i+1}, \delta'_{i,j}) + \mathcal{P}(\lambda'_{i,j} E_i),$$

where the right-hand side should be interpreted as the sum of two independent random variables, and with

$$\begin{aligned} \delta'_{i,j} &:= \prod_{k=n-i+1}^{j-1} (1 - \delta_k), \\ \lambda'_{i,j} &:= \sum_{k=n-i+2}^j \lambda_k \left(\prod_{\ell=k}^{j-1} (1 - \delta_\ell) \right). \end{aligned}$$

Proof. We prove the lemma by induction. Let $1 \leq i \leq n$ be fixed. Since:

$$C_{i,n-i+2} = C_{i,n-i+1} - D_{i,n-i+2} + N_{i,n-i+2},$$

from Lemma 3.3, $C_{i,n-i+1} - D_{i,n-i+2} \mid \mathcal{F}_n \sim \mathcal{B}(C_{i,n-i+1}, 1 - \delta_{i,n-i+1})$ and $N_{i,n-i+2} \mid \mathcal{F}_n \sim \mathcal{P}(\lambda_{n-i+2} E_i)$. Assume by induction that:

$$C_{i,j} \mid \mathcal{F}_n \sim \mathcal{B}(C_{i,n-i+1}, \delta'_{i,j}) + \mathcal{P}(\lambda'_{i,j} E_i).$$

Remark that

$$C_{i,j+1} = C_{i,n-i+1} + (C_{i,j} - C_{i,n-i+1}) + N_{i,j+1} - D_{i,j+1}.$$

and $D_{i,j+1} \sim \mathcal{B}(C_{i,j}, \delta_j) = \mathcal{B}(C_{i,n-i+1}, \delta_j) + \mathcal{B}(C_{i,j} - C_{i,n-i+1}, \delta_j)$, where the right-hand side should be interpreted as the sum of two independent random variables.

Under $H2'$, by Lemma 3.3,

$$C_{i,n-i+1} - \mathcal{B}(C_{i,n-i+1}, \delta_j) \mid \mathcal{F}_n \sim \mathcal{B}(C_{i,n-i+1}, \delta'_{i,j+1}),$$

and

$$(C_{i,j} - C_{i,n-i+1}) - \mathcal{B}(C_{i,j} - C_{i,n-i+1}, \delta_j) \mid \mathcal{F}_n \sim \mathcal{P}(\lambda'_{i,j} E_i (1 - \delta_j))$$

and finally, by Lemma 3.1

$$C_{i,j+1} \mid \mathcal{F}_n \sim \mathcal{B}(C_{i,n-i+1}, \delta'_{i,j+1}) + \mathcal{P}(\lambda'_{i,j+1} E_i).$$

□

The result indicates that, following the observation of \mathcal{F}_n , the distribution of an unobserved $C_{i,j}$ can be described as the sum of two components: current claims that exceed the priority threshold and are likely to stay above it, and new claims that may initially rise above the priority but might later fall below it.

If we are interested in estimating $\mathbb{E}(C_{i,n} \mid \mathcal{F}_n)$, we can use the estimator $\widehat{C}_{i,n} := \widehat{\lambda}'_{i,n} E_i + \widehat{\delta}'_{i,n} C_{i,n-i+1}$ where $(\widehat{\lambda}'_{i,n}, \widehat{\delta}'_{i,n})$ are estimators for $(\lambda'_{i,n}, \delta'_{i,n})$. The sequences (λ_k) and (δ_k) are estimated by $(\widehat{\lambda}_k)$ and $(\widehat{\delta}_k)$, as defined in (2).

In practice, the Poisson distribution is commonly used for modeling the number of claims due to its simplicity and the assumption of independent arrivals with non-random intensity across all claims. However, it is often observed that the empirical variance exceeds the empirical mean, suggesting that the claim data might not fully adhere to the assumptions of the Poisson distribution.

For instance, if the intensity parameter of claim arrivals for each policy follows a Gamma distribution, the total number of claims is known to follow a Negative binomial distribution. This distribution is favored because it remains straightforward to use, accommodates excess variance, and converges to the Poisson distribution when this variance diminishes.

In the following subsection, we propose using the Negative binomial distribution instead of the Poisson, demonstrating that, with an appropriate assumption, it can yield comparable results.

3.2 The Negative binomial assumption

In this section, we aim to establish a Negative binomial framework that yields results similar to those obtained from the previous Poisson framework. Specifically, we define the Negative binomial distribution as follows: let $N \sim \mathcal{NB}(r, p)$ where $r > 0$ and $0 < p < 1$. The probability mass function of N is given by

$$\mathbb{P}(N = n) = \frac{\Gamma(r + n)}{n! \Gamma(r)} p^r (1 - p)^n, \quad n \in \mathbb{N}. \quad (3)$$

We then proceed by introducing a new assumption that serves as a replacement for $H2'$. The change of this assumption has an impact only on the sequence of random variables $(N_{i,j})_{1 \leq i, j \leq n}$.

Assumption 3.2.

$H2''$ For $1 \leq j \leq n$, there exists $r_j \geq 0$ and for $1 \leq j \leq n - 1$, there exists $0 \leq \delta_j \leq 1$ such that

$$\begin{aligned} N_{i,j} \mid \mathcal{F}_{i+j-2} &\sim \mathcal{NB}(r_j E_i, p_j) \text{ in which } p_{j+1} := \frac{p_j}{1 - \delta_j(1 - p_j)}, \quad 1 \leq i \leq n, \\ D_{i,j+1} \mid \mathcal{F}_{i+j-1} &\sim \mathcal{B}(C_{i,j}, \delta_j), \quad 1 \leq i \leq n. \end{aligned}$$

Similarly, $H2''$ implies $H2$ and $H3$ from Section 2. The structure of the p 's may not initially appear clear or intuitive. However, the representation of these parameters will be clarified later. The free parameters are the $(r_j)_{1 \leq j \leq n}$ (which replace the λ 's from $H2'$), the $(\delta_j)_{1 \leq j \leq n-1}$ and $p_1 \in]0, 1[$. There is only one additional parameter, compared to $H2'$. This extra parameter governs the additional variance due to the specific configuration of the family $(p_j)_{1 \leq j \leq n}$.

Remark 3.5. The p 's can be explicitly expressed in terms of p_1 and the δ 's:

$$p_j = \frac{p_1}{1 - (1 - p_1) \left[\sum_{k=1}^{j-1} \delta_k \prod_{\ell=1}^{k-1} (1 - \delta_\ell) \right]}, \quad 1 \leq j \leq n.$$

The above remark can be verified through direct induction. Estimating the r 's, δ 's, and p_1 cannot yield explicit maximum likelihood estimates. For $\hat{\delta}$'s, we can use the same estimator as the one defined in the Poisson framework in (2). When p_1 is known, to estimate the r 's, we can use a moment estimator based on the expected value and set:

$$\hat{r}_j := \hat{\lambda}_j \frac{p_j}{1 - p_j}, \quad 1 \leq j \leq n. \quad (4)$$

Finally, the estimation of p_1 can be computed numerically using maximum likelihood methods:

$$\hat{p}_1 \in \arg \max \sum_{j=1}^n \sum_{i=1}^{n-j+1} \log f(N_{i,j}, \hat{r}_j E_i, p_j), \quad (5)$$

in which $(n, r, p) \mapsto f(n, r, p)$ denotes the probability mass function of the Negative binomial distribution with parameter (r, p) , see (3).

It remains to explain why we choose this form for the p 's. It is a consequence of the following result, in order to have a consistent form, as in the Poisson case. To establish the distribution of the $(C_{i,j})$, we begin with a classic lemma.

Lemma 3.6. *Let N be a random variable with distribution $\mathcal{NB}(r, p)$ for $r > 0$ and $0 < p < 1$, and D a random variable such that $D \mid N \sim \mathcal{B}(N, \delta)$ for $0 < \delta < 1$. Then*

$$N - D \sim \mathcal{NB}(r, p') \quad \text{with} \quad p' := \frac{p}{1 - \delta(1 - p)}.$$

Proof. Let $k \in \mathbb{N}$.

$$\begin{aligned} \mathbb{P}(N - D = k) &= \sum_{n \in \mathbb{N}} \mathbb{P}(\{D = n - k\} \mid \{N = n\}) \mathbb{P}(N = n) \\ &= \sum_{n \geq k} \frac{n!}{k!(n-k)!} \delta^{n-k} (1-\delta)^k \frac{\Gamma(r+n)}{n! \Gamma(r)} p^r (1-p)^n \\ &= \frac{(1-\delta)^k}{k!} p^r \sum_{n \geq 0} \frac{\delta^n \Gamma(r+n+k)}{n! \Gamma(r)} (1-p)^{n+k} \\ &= \frac{[(1-\delta)(1-p)]^k}{k!} p^r \sum_{n \geq 0} \frac{\Gamma(r+n+k)}{n! \Gamma(r)} [\delta(1-p)]^n \\ &= \frac{\Gamma(r+k)}{k! \Gamma(r)} \left[\frac{(1-\delta)(1-p)}{1-\delta(1-p)} \right]^k \left[\frac{p}{1-\delta(1-p)} \right]^r. \end{aligned}$$

The final line is obtained by noting that $\sum_{n \geq 0} \frac{\Gamma(r+k+n)}{n! \Gamma(r+k)} [1-\delta(1-p)]^{r+k} [\delta(1-p)]^n = 1$. \square

Proposition 3.7. *For all $1 \leq i, j \leq n$, we have*

$$C_{i,j} \sim \mathcal{NB}(r'_j E_i, p_j),$$

with

$$r'_j := \sum_{k=1}^j r_k.$$

Proof. We prove the lemma by induction. Let $1 \leq i \leq n$ be fixed. By construction $C_{i,1} = N_{i,1} \sim \mathcal{NB}(r_1 E_i, p_1)$. Assume by induction that $C_{i,j}$ follows a Negative binomial distribution with parameter $(r'_j E_i, p_j)$. Recall that

$$C_{i,j+1} = C_{i,j} + N_{i,j+1} - D_{i,j+1}.$$

Under $H2''$, by Lemma 3.6,

$$C_{i,j} - D_{i,j+1} \sim \mathcal{NB}(r'_j E_i, p_{j+1}),$$

and finally:

$$C_{i,j+1} \sim \mathcal{NB}(r'_{j+1} E_i, p_{j+1}).$$

\square

The form of the p 's can now be understood. Assuming new claims above the priority threshold follow a Negative binomial distribution and that some claims may later fall below the priority, we aim to maintain consistency at any point in time with the Negative binomial distribution. Consequently, this leads to the specific form of the p 's.

The p 's are increasing, leading to a smaller excess of variance over time. Notably, the likelihood of claims dropping below the priority with probabilities δ 's influences the Negative binomial distribution, including claims not yet reported in future development years. And more likely the claim are dropping below the threshold, faster decreases the excess of variance for the new claims.

In particular, for the extremal cases, if $\delta_j = 0$, meaning that no claims drops below the priority, the excess of variance does not reduce. Conversely, when $\delta_j \rightarrow 1$, meaning all claims drop below the priority, the excess variance vanishes.

The preceding result provides the distribution of $C_{n+1,n}$ given the related exposure. Additionally, we may be interested in the distribution of $C_{i,j} | \mathcal{F}_n$ for $i + j > n + 1$, as described in the following proposition.

Proposition 3.8. *For all $1 \leq i, j \leq n$ such that $i + j > n + 1$, we have*

$$C_{i,j} | \mathcal{F}_n \sim \mathcal{B}(C_{i,n-i+1}, \delta'_{i,j}) + \mathcal{NB}(r'_j E_i, p_j),$$

where the right-hand side should be interpreted as the sum of two independent random variables, and with

$$\delta'_{i,j} := \prod_{k=n-i+1}^{j-1} (1 - \delta_k),$$

$$r'_j := \sum_{k=n-i+2}^j r_k.$$

Proof. The proof follows the same reasoning as in Proposition 3.4, with the key difference being the application of Lemma 3.6 in place of Lemma 3.1. \square

If we are interested in estimating $\mathbb{E}(C_{i,n} | \mathcal{F}_n)$, using (4) for the estimators of the r 's leads to the same estimator as the one suggested for Proposition 3.4.

4 Bootstrap methodology

In [2], the author discusses a bootstrap methodology for the general Schnieper model to resimulate the λ 's and δ 's, accounting for uncertainty in the parameters. They also simulate a Gaussian random variable to integrate the internal randomness of the process for each development stage. This follows the main ideas of the non-parametric bootstrap as summarized in [1].

Here, we present a distinct approach that utilizes the specific framework of claim numbers and proposes a comprehensive parametric bootstrap methodology without computing residuals or making any additional assumption.

4.1 The Poisson case

Let $M \in \mathbb{N}^*$ be the total number of bootstrap simulations to be performed. To account for the inherent randomness of the λ 's and δ 's, for $1 \leq m \leq M$. To do it efficiently, we shall use the following lemma.

Lemma 4.1. *Under $H1$ and $H2'$, we have:*

$$\widehat{\lambda}_j | \mathcal{B}_{j-1} \sim \frac{\mathcal{P}\left(\lambda_j \sum_{i=1}^{n-j+1} E_i\right)}{\sum_{i=1}^{n-j+1} E_i}, \quad 1 \leq j \leq n,$$

$$\widehat{\delta}_j | \mathcal{B}_j \sim \frac{\mathcal{B}\left(\sum_{i=1}^{n-j} C_{i,j}, \delta_j\right)}{\sum_{i=1}^{n-j} C_{i,j}}, \quad 1 \leq j \leq n-1,$$

where $\mathcal{B}_k := \sigma(N_{i,j}, D_{i,j} \mid i+j \leq n+1, j \leq k)$.

Proof. Direct consequence of $H1$ and $H2'$. □

This provides a direct method to simulate the bootstrapped λ 's and δ 's.

$$(\widehat{\lambda}_j^m)_{1 \leq m \leq M} \stackrel{i.i.d.}{\sim} \frac{\mathcal{P}\left(\widehat{\lambda}_j \sum_{i=1}^{n-j+1} E_i\right)}{\sum_{i=1}^{n-j+1} E_i}, \quad 1 \leq j \leq n,$$

$$(\widehat{\delta}_j^m)_{1 \leq m \leq M} \stackrel{i.i.d.}{\sim} \frac{\mathcal{B}\left(\sum_{i=1}^{n-j} C_{i,j}, \widehat{\delta}_j\right)}{\sum_{i=1}^{n-j} C_{i,j}}, \quad 1 \leq j \leq n-1,$$

where $(\widehat{\lambda}_j)_{1 \leq j \leq n}$ and $(\widehat{\delta}_j)_{1 \leq j \leq n-1}$ come from (2) and the C 's are the observed data.

Bootstrap simulation of $C_{n+1,n}$. Following Proposition 3.2, the bootstrap simulation is performed as follows:

$$C_{n+1,n}^m \sim \mathcal{P}(\widehat{\lambda}_n^m E_{n+1}), \quad 1 \leq m \leq M.$$

Bootstrap simulation of $C_{i,j} \mid \mathcal{F}_n$. For the lower triangle, the simulation is conducted using:

$$\begin{aligned} C_{i,n-i+1}^m &:= C_{i,n-i+1}, & 1 \leq i \leq n, \\ C_{i,j+1}^m &:= C_{i,j}^m + \mathcal{P}\left(\widehat{\lambda}_{j+1}^m E_i\right) - \mathcal{B}\left(C_{i,j}^m, \widehat{\delta}_j^m\right), & n-i+1 \leq j \leq n-1. \end{aligned}$$

On the right-hand side of the last equality, the difference of the two distributions should be interpreted as the difference of two independent random variables. This procedure generates a bootstrap distribution for the random variable $C_{i,j} \mid \mathcal{F}_n$ on the lower triangle. In this process, the uncertainty associated with the estimators of the parameters is integrated.

Remark 4.2. Based on Proposition 3.4, when our focus is on $C_{i,n} \mid \mathcal{F}_n$, we can efficiently simulate $C_{i,n}^m$, using

$$C_{i,n}^m \sim \mathcal{B}(C_{i,n-i+1}, \widehat{\delta}_{i,n}^m) + \mathcal{P}(\widehat{\lambda}_{i,n}^m E_i),$$

where the right-hand side should be interpreted as the sum of two independent random variables, and with

$$\begin{aligned} \widehat{\delta}_{i,n}^m &:= \prod_{k=n-i+1}^{n-1} (1 - \widehat{\delta}_k^m), \\ \widehat{\lambda}_{i,n}^m &:= \sum_{k=n-i+2}^n \widehat{\lambda}_k^m \left(\prod_{\ell=k}^{n-1} (1 - \widehat{\delta}_\ell^m) \right). \end{aligned}$$

This provides a more efficient algorithm.

4.2 The Negative binomial case

We extend the Poisson model to fit the Negative binomial framework. To account for the variability in the parameters r 's, δ 's and p_1 , for $1 \leq m \leq M$. We have the following lemma, similar to Lemma 4.1.

Lemma 4.3. Under $H1$ and $H2''$, we have:

$$\begin{aligned} \widehat{\lambda}_j \mid \mathcal{B}_{j-1} &\sim \frac{\mathcal{NB}\left(r_j \sum_{i=1}^{n-j+1} E_i, p_j\right)}{\sum_{i=1}^{n-j+1} E_i}, & 1 \leq j \leq n, \\ \widehat{\delta}_j \mid \mathcal{B}_j &\sim \frac{\mathcal{B}\left(\sum_{i=1}^{n-j} C_{i,j}, \delta_j\right)}{\sum_{i=1}^{n-j} C_{i,j}}, & 1 \leq j \leq n-1, \end{aligned}$$

where $\mathcal{B}_k := \sigma(N_{i,j}, D_{i,j} \mid i+j \leq n+1, j \leq k)$.

Proof. Direct consequence of $H1$ and $H2''$. □

However, unlike the Poisson case, we cannot apply the above lemma straightforwardly since it does not provide the distribution of p_1 . Additionally, the estimator of p_1 , defined in (5), is non-trivially dependent on the $N_{i,j}$.

Nonetheless, for the δ 's, we can proceed as follows:

$$(\widehat{\delta}_j^m)_{1 \leq m \leq M} \stackrel{i.i.d.}{\sim} \frac{\mathcal{B}\left(\sum_{i=1}^{n-j} C_{i,j}, \widehat{\delta}_j\right)}{\sum_{i=1}^{n-j} C_{i,j}}, \quad 1 \leq j \leq n-1.$$

For the r 's and p_1 , we resimulate the upper triangle $N_{i,j}$.

$$N_{i,j}^m \sim \mathcal{NB}(\widehat{r}_j E_i, \widehat{p}_j), \quad 1 \leq i \leq n, \quad 1 \leq j \leq n-i.$$

From these upper triangles, we can estimate the r 's and p_1 . These are denoted respectively as $(\widehat{r}_j^m)_{1 \leq j \leq n}$ and \widehat{p}_1^m .

Bootstrap simulation of $C_{n+1,n}$. Following Proposition 3.7, the simulation is straightforward:

$$C_{n+1,n}^m \sim \mathcal{NB}(\widehat{r}_n^m E_{n+1}, \widehat{p}_n^m), \quad 1 \leq m \leq M.$$

Bootstrap simulation of $C_{i,j} \mid \mathcal{F}_n$. For the lower triangle, we simulate :

$$\begin{aligned} C_{i,n-i+1}^m &:= C_{i,n-i+1}, & 1 \leq i \leq n, \\ C_{i,j+1}^m &:= C_{i,j}^m + \mathcal{NB}(\widehat{r}_{j+1}^m E_i, \widehat{p}_{j+1}^m) - \mathcal{B}(C_{i,j}^m, \widehat{\delta}_j^m), & n-i+1 \leq j \leq n-1. \end{aligned}$$

On the right-hand side of the last equality, the difference of the two distributions should be interpreted as the difference of two independent random variables. Similarly to the approach discussed in Remark 4.2 for the Poisson case, if our focus is solely on the distribution $C_{i,n} \mid \mathcal{F}_n$, we can bypass simulating the entire lower triangle using Proposition 3.8.

5 Example

We present two triangles of simulated data to illustrate both cases with $n = 6$ years of observations. For the first one, the exposure and the C triangle are:

i	E_i	$C_{i,1}$	$C_{i,2}$	$C_{i,3}$	$C_{i,4}$	$C_{i,5}$	$C_{i,6}$
1	20	5	9	11	12	13	11
2	25	11	16	13	11	17	
3	32	9	17	22	22		
4	38	10	10	11			
5	42	17	18				
6	45	14					

whose decomposition in N and D is:

i	$N_{i,1}$	$N_{i,2}$	$N_{i,3}$	$N_{i,4}$	$N_{i,5}$	$N_{i,6}$	i	$D_{i,1}$	$D_{i,2}$	$D_{i,3}$	$D_{i,4}$	$D_{i,5}$	$D_{i,6}$
1	5	4	5	2	1	0	1	0	0	3	1	0	2
2	11	9	4	4	6		2	0	4	7	6	0	
3	9	14	9	3			3	0	6	4	3		
4	10	7	5				4	0	7	4			
5	17	10					5	0	9				
6	14						6	0					

We can derive directly the λ 's and δ 's.

j	1	2	3	4	5	6
$\widehat{\lambda}_j$	0.327	0.28	0.2	0.117	0.156	0
$\widehat{\delta}_j$	0.5	0.346	0.217	0	0.154	–

Let $E_{n+1} = 50$ be the exposure for the upcoming year. Under the Poisson assumption, using the estimator of the intensity leads to:

$$C_{n+1,n} \sim \mathcal{P}(27.752)$$

Under the Negative binomial assumption and utilizing the λ 's and the δ 's, and computing p_1 by maximum likelihood leads to optimal $p_1 \rightarrow 1$. In this case, the Negative binomial distribution converges to the Poisson distribution: the assumption does not appear suitable.

To account for the uncertainty in the unknown parameters, we can use the bootstrap procedure. We get that the variance of $C_{n+1,n}$ is now around 53.361, which is notably higher than the variance obtained when using the Poisson distribution with the estimated parameter. Figure 1 illustrates the histogram of $C_{n+1,n}$ with the distribution $\mathcal{P}(27.752)$ (lighter on the left) compared to the distribution obtained from the bootstrap procedure (darker on the right).

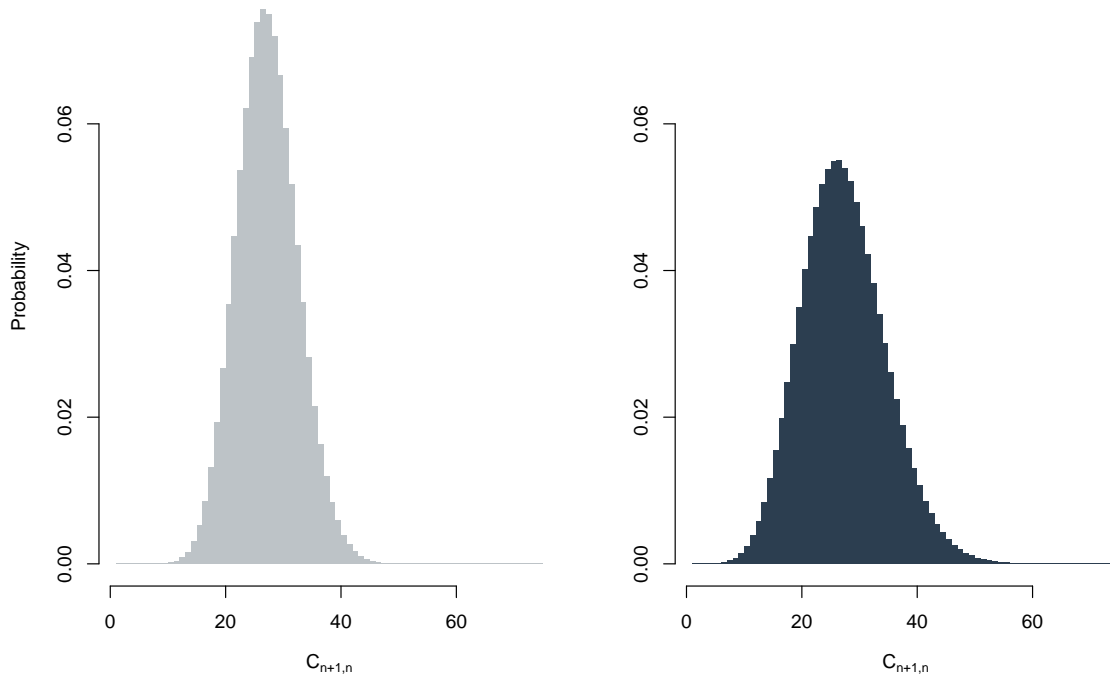


Figure 1: Comparison of the distribution of the estimated distribution $\mathcal{P}(27.752)$ (on the left), and the associated bootstrap distribution (on the right); with $M = 10^7$ simulations.

We now introduce a second set of triangles, the exposure and the C triangle are:

i	E_i	$C_{i,1}$	$C_{i,2}$	$C_{i,3}$	$C_{i,4}$	$C_{i,5}$	$C_{i,6}$
1	20	8	4	12	12	14	13
2	25	3	5	7	10	15	
3	32	5	10	11	9		
4	38	27	20	29			
5	42	23	18				
6	45	14					

whose decomposition in N and D is:

i	$N_{i,1}$	$N_{i,2}$	$N_{i,3}$	$N_{i,4}$	$N_{i,5}$	$N_{i,6}$
1	8	3	9	4	3	0
2	3	5	4	3	6	
3	5	7	3	3		
4	27	8	13			
5	23	7				
6	14					

i	$D_{i,1}$	$D_{i,2}$	$D_{i,3}$	$D_{i,4}$	$D_{i,5}$	$D_{i,6}$
1	0	7	1	4	1	1
2	0	3	2	0	1	
3	0	2	2	5		
4	0	15	4			
5	0	12				
6	0					

Again, we can derive directly the λ 's and δ 's.

j	1	2	3	4	5	6
$\widehat{\lambda}_j$	0.396	0.191	0.252	0.13	0.2	0
$\widehat{\delta}_j$	0.591	0.231	0.3	0.091	0.071	-

Given an exposure of $E_{n+1} = 50$ for the next year and assuming a Poisson distribution, we get that

$$C_{n+1,n} \sim \mathcal{P}(30.243)$$

Computing p_1 by maximum likelihood does not lead to $p_1 \rightarrow 1$ anymore. We find $\widehat{p}_1 = 0.397$. This implies that

j	1	2	3	4	5	6
\widehat{p}_j	0.397	0.616	0.676	0.749	0.767	0.78
\widehat{r}_j	0.263	0.402	0.418	0.448	0.328	0.177

In particular, under the Negative binomial assumption,

$$C_{n+1,n} \sim \mathcal{NB}(106.94, 0.780).$$

By construction, the expected value of $C_{n+1,n}$ remains at 30.243, but the variance increases to 38.796.

To choose between the two assumptions, we can calculate the log-likelihood and AIC for both cases. Table 1 presents the results.

	log-L.	AIC
\mathcal{P}	-53.937	119.875
\mathcal{NB}	-50.793	115.586

Table 1: Log-likelihood and AIC for both assumptions.

It appears that using the Negative binomial distribution is the most suitable choice in this scenario. Given that the δ 's are also part of the definition of the p 's, one might question whether it would be beneficial to estimate both $(p_1, (\delta_j)_{1 \leq j \leq n-1})$ simultaneously using data from both N and D . Let $(\tilde{p}_1, (\tilde{\delta}_j)_{1 \leq j \leq n-1})$ denote the new estimators. Table 2 provides a comparison, which shows that the difference is minimal.

j	1	2	3	4	5
$\hat{\delta}_j$	0.591	0.231	0.3	0.091	0.071
$\tilde{\delta}_j$	0.601	0.232	0.305	0.092	0.071

j	1	2	3	4	5	6
\hat{p}_j	0.397	0.616	0.676	0.749	0.767	0.78
\tilde{p}_j	0.393	0.619	0.679	0.753	0.77	0.783

Table 2: Comparison of the two estimation methods.

Figure 2 shows the distribution from the Poisson assumption (lighter on the left) compared with the distribution from the negative binomial assumption (darker on the right).

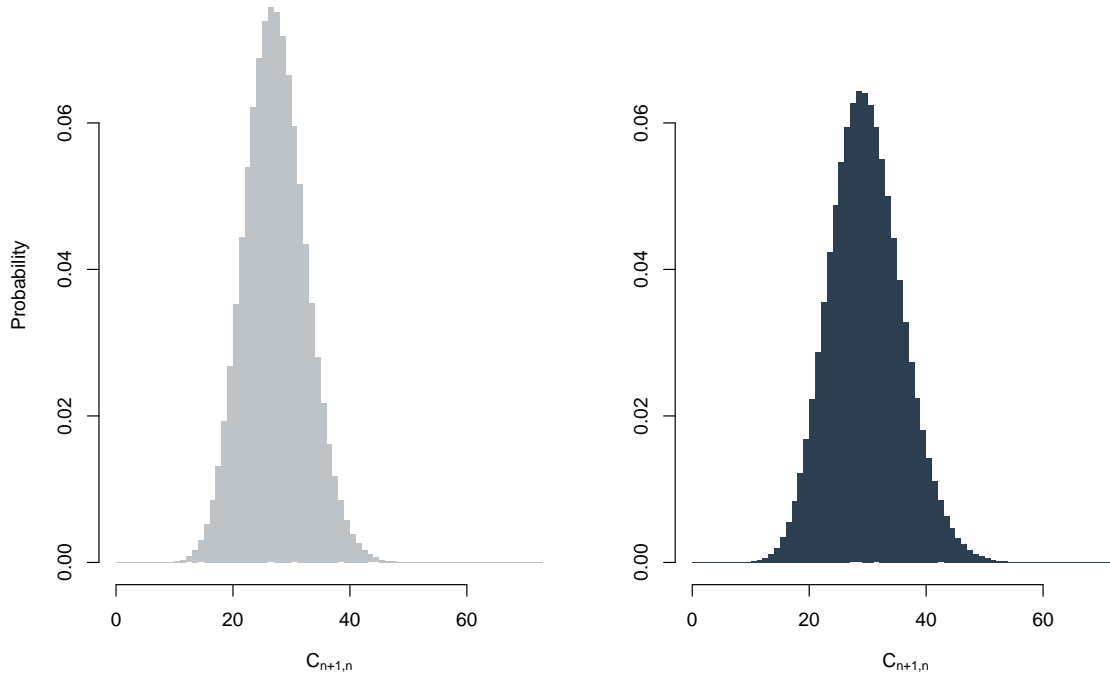


Figure 2: Comparison of the distribution of the estimated distribution $\mathcal{P}(30.243)$ (on the left), and the distribution of the estimated distribution $\mathcal{NB}(106.939, 0.780)$ (on the right).

Figure 3 illustrates the bootstrap distribution from the Poisson assumption (lighter on the left) and from the Binomial negative assumption (darker on the right). In each simulation using the negative binomial approach, if the estimated \hat{p}_1 was close to 1, suggesting that the Poisson distribution was a better fit, the simulation was conducted using the Poisson framework. The bootstrap results show that the variance of the distribution under the Poisson assumption is 62.633, while the variance under the negative binomial assumption is 67.658.

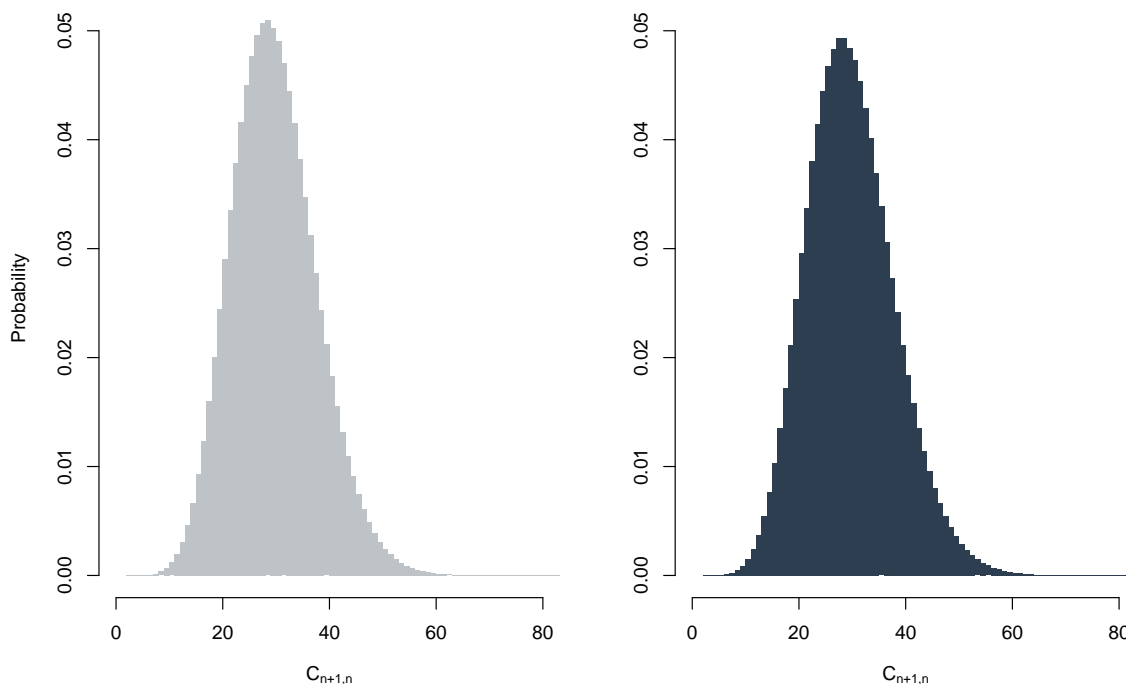


Figure 3: Comparison of the Bootstrap distribution obtained with the Poisson assumption (on the left), and with the Negative binomial assumption (on the right); with $M = 10^7$ simulations.

Acknowledgments

The author acknowledges the financial support provided by the *Fondation Natixis*.

References

- [1] P. D. England and R. J. Verrall. Predictive distributions of outstanding liabilities in general insurance. *Annals of Actuarial Science*, 1(2):221–270, 2006.
- [2] Huijuan Liu and Richard Verrall. A bootstrap estimate of the predictive distribution of outstanding claims for the schnieper model. *ASTIN Bulletin*, 39(2):677–689, 2009.
- [3] Huijuan Liu and Richard Verrall. Predictive distributions for reserves which separate true ibnr and ibner claims. *ASTIN Bulletin*, 39(1):35–60, 2009.
- [4] Thomas Mack. Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, 23(2):213–225, 1993.
- [5] Esbjörn Ohlsson. Using separate exposure for ibnyr and ibner in the chain ladder method, 2015.

- [6] Esbjörn Ohlsson and Björn Wallberg-Beutelrock. Claims reserving using separate exposure for claims with and without a case reserve, 2022.
- [7] R. Schnieper. Separating true ibnr and ibner claims. *ASTIN Bulletin*, 21(1):111–127, 1991.