IRM | institut du risk management BY Institut.

Mémoire présenté pour la validation de la Formation
« Certificat d'Expertise Actuarielle »
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le

Par : Ao SUN

Titre : Approche bayésienne pour la tarification a posteriori en assurance santé collective

Confidentialité : ☒ NON ☐ OUI (Durée : ☐ 1an ☐ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

_____

_____

_____

Membres présents du jury de l'Institut du Risk Management :

_____

_____

_____

_____

_____

_____

_____

_____

Secrétariat :

Bibliothèque :

Entreprise : Malakoff Humanis

Nom : _____

Signature et Cachet :

Directeur de mémoire en entreprise :

Nom : Nathalie BRESSON PALEY

Signature :

Invité :

Nom : _____

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels
(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise

Signature(s) du candidat(s)

Résumé

Le secteur de l'assurance santé collective est un domaine dynamique, confronté à des changements réglementaires significatifs, tels que la généralisation de la couverture complémentaire santé et la réforme du 100% santé. Ces évolutions, couplées à la possibilité de résiliation infra-annuelle, accentuent la concurrence sur le marché. Dans ce contexte, il est crucial pour les assureurs de maîtriser les risques et d'adapter leurs primes de manière réactive et stratégique.

Ce mémoire présente une approche bayésienne avancée pour la tarification a posteriori en assurance santé collective, exploitant des données riches et diversifiées pour modéliser avec précision les différentes garanties de santé. L'application de méthodes statistiques avancées permet de saisir la complexité et la variabilité inhérentes à ces données.

Des exemples concrets illustrent l'utilisation de cette méthodologie bayésienne pour élaborer des primes pures ajustées. Cette stratégie permet une tarification flexible, adaptée aux spécificités de chaque entreprise.

Le mémoire se penche également sur l'utilisation des modèles hiérarchiques bayésiens. Ces modèles, en structurant les données de manière hiérarchique, fournissent une compréhension nuancée des risques et facilitent une segmentation fine du marché. Cette approche novatrice ouvre des perspectives enrichissantes pour la tarification actuarielle, en alignant les modèles tarifaires sur les réalités complexes du secteur de la santé collective.

Mots-clés : *complémentaire santé*, *tarification a posteriori*, *modèle bayésien*, *modèle de mélange*, *modèle hiérarchique bayésien.*

Abstract

The group health insurance sector is a dynamic field, facing significant regulatory changes, such as the generalization of supplementary health coverage and the « 100% Santé » reform. These developments, combined with the possibility of intra-annual termination, intensify market competition. In this context, it is crucial for insurers to manage risks and adapt their pricing strategies responsively and strategically.

This thesis presents an advanced bayesian approach to a posteriori pricing in group health insurance, leveraging rich and diverse data to accurately model various health coverage guarantees. The application of sophisticated statistical methods captures the inherent complexity and variability of these data.

Concrete examples illustrate the use of this bayesian methodology in developing adjusted insurance premiums. This strategy allows for flexible pricing, tailored to the specificities of each company.

The thesis also delves into the use of bayesian hierarchical models. These models, by structuring data hierarchically, provide a nuanced understanding of risks and facilitate a fine segmentation of the market. This innovative approach opens up enriching perspectives for actuarial pricing, aligning pricing models with the complex realities of the group health sector.

Keywords: *complementary health insurance*, *a posteriori pricing*, *bayesian model*, *mixture model*, *bayesian hierarchical model*.

Note de Synthèse

Ce mémoire se penche sur les méthodes de tarification dans le domaine de l'assurance santé collective, en mettant un accent particulier sur l'approche bayésienne pour la tarification a posteriori. S'appuyant sur un ensemble volumineux de données historiques réelles, cette étude démontre, à travers des cas d'étude concrets, l'efficacité et la pertinence de l'utilisation de modèles bayésiens dans la tarification a posteriori pour l'assurance santé collective. Elle fournit une analyse détaillée de la façon dont ces modèles peuvent être mobilisés pour relever les enjeux spécifiques de ce secteur, offrant ainsi des perspectives novatrices et des solutions avancées pour la complexité inhérente à la tarification actuarielle.

Problématique et motivation :

Le secteur de l'assurance santé collective, caractérisé par sa dynamique constante, fait face à des défis réglementaires significatifs et à une concurrence accrue suite à l'introduction de la résiliation infra-annuelle. Dans cet environnement compétitif, il est essentiel pour les assureurs de maîtriser efficacement les risques et d'ajuster leurs primes de manière réactive et stratégiquement affinée.

La tarification a posteriori, qui représente une approche dynamique ajustant les primes en fonction de l'expérience de sinistralité réelle des assurés, se révèle cruciale pour maintenir la rentabilité et répondre aux attentes financières et stratégiques dans la conception des produits d'assurance.

Cependant, les méthodes traditionnelles de tarification a posteriori tendent à nécessiter une agrégation de données à un niveau relativement général. De plus, les approches modernes issues de la "Data Science", telles que les modèles linéaires généralisés ou les méthodes fondées sur les arbres de décision, imposent des structures de données strictes et présentent des limites, notamment en cas d'absence de sinistres passés.

L'approche bayésienne se présente alors comme une solution prometteuse. Sa capacité à intégrer l'expertise à travers les distributions a priori avant même l'analyse des données, associée à sa flexibilité, son interprétabilité, et sa prise en compte efficace des effets d'absence de sinistres, en fait un choix stratégique avantageux.

Historiquement, l'application de l'approche bayésienne a été limitée par sa complexité de calcul. Cependant, grâce aux avancées méthodologiques telles que les méthodes de Monte Carlo par chaînes de Markov (MCMC) et aux progrès technologiques, y compris le développement de logiciels tels que PyStan et PyMC, cette approche est devenue plus accessible. Dans ce contexte, nous avons choisi d'appliquer l'approche bayésienne aux données volumineuses pour la tarification a posteriori, ouvrant ainsi la voie à des solutions innovantes dans le domaine de l'assurance santé collective.

Méthodologie suivie :

La tarification a posteriori en assurance santé collective, abordée dans ce mémoire, repose sur une approche méthodologique qui s'est articulée autour de plusieurs étapes clés, visant à assurer une

analyse précise et pertinente des données dans le cadre de l'application de modèles bayésiens. Voici les principales composantes de cette méthodologie :

- **Préparation des données et hypothèses de travail** : La méthodologie de ce mémoire a débuté par une préparation exhaustive des données, un processus jugé essentiel pour assurer la qualité et la pertinence des informations en vue de la modélisation. Cette étape a impliqué la collecte, l'analyse, le nettoyage et l'agrégation des données issues de diverses sources. Il a été crucial de reconnaître et de gérer les imperfections des données brutes, notamment les valeurs anormales ou aberrantes, et de les agréger judicieusement pour maintenir un équilibre entre la généralisation des modèles et la conservation des informations essentielles.

- **Analyse descriptive des données** : L'analyse descriptive des données a constitué une étape fondamentale, offrant un aperçu complet des distributions statistiques des données, y compris leurs valeurs minimales et maximales. Cette analyse a permis d'évaluer la forme générale et les tendances des données, et de choisir le type de distribution le plus approprié pour la loi a priori dans les modèles bayésiens, un choix crucial pour la précision et la validité des modèles prédictifs.

- **Application des modèles bayésiens** : Après une préparation et une analyse approfondies des données, l'étape suivante a été l'application de modèles bayésiens. Cette approche a permis d'exploiter les nuances et les modèles cachés dans les données en vue d'une tarification a posteriori précise. Cette étape a représenté un pont entre la préparation minutieuse des données et leur application pratique dans le monde complexe de l'actuariat, démontrant la puissance de l'analyse actuarielle soutenue par des méthodes de modélisation avancées.

Travaux réalisés et interprétation des résultats :

Initialement, notre étude a exploré le contexte de l'assurance santé collective, en se concentrant sur les réglementations et les réformes récentes dans ce domaine. Nous avons scruté attentivement les rapports et les statistiques globales du marché de l'assurance santé, afin d'acquérir une compréhension approfondie du contexte d'étude.

Par la suite, notre attention s'est portée sur l'analyse des données du groupe Malakoff Humanis, en vue d'examiner l'utilisation de l'approche bayésienne dans la tarification a posteriori. Pour cela, il était essentiel de définir un périmètre d'étude spécifique, doté de données complètes et de haute qualité.

Après avoir délimité le périmètre de l'étude, nous nous sommes lancés dans la collecte et l'analyse des données issues de diverses sources. Cette phase, bien que chronophage et exigeante, est essentielle pour garantir la qualité de la modélisation. Dans le cadre de cette étude, plus de 10 millions de lignes de données ont été collectées et analysées.

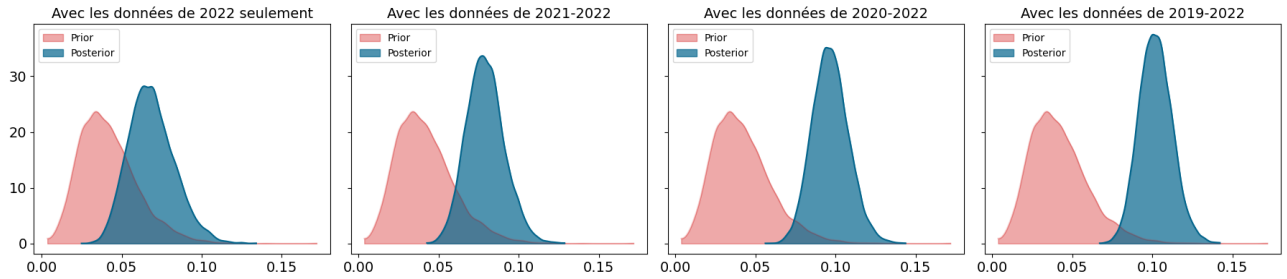
Un accent particulier a été mis sur le regroupement des actes de soins. Grâce à une analyse des libellés et des critères tels que la fréquence et les montants, nous avons pu former environ quarante catégories de regroupement, facilitant les analyses descriptives et réduisant la taille des données pour une étude plus précise et maniable.

L'étape de calibration des données, en fonction de la présence proportionnelle des bénéficiaires, s'est révélée cruciale pour une analyse exacte, surtout en raison des variations de la durée de couverture au

sein des contrats collectifs de santé complémentaire.

Pour chaque regroupement d'actes, deux modèles ont été développés : l'un pour estimer le nombre d'occurrences annuelles et l'autre pour les montants de remboursement associés. Comme illustré dans le graphique 1 ci-dessous, la distribution a posteriori, qui intègre les données historiques, diverge progressivement de la distribution a priori, reflétant ainsi les effets individuels.

FIGURE 1 – Évolution de la distribution de paramètre avec l'intégration des données



En utilisant les distributions a posteriori estimées pour un regroupement d'actes de soins, nous avons simulé le nombre d'occurrences de soins et les montants de remboursement associés. Le produit de ces deux éléments nous a permis d'estimer les coûts envisagés pour l'assureur, et donc la prime pure. En additionnant les primes pures a posteriori pour tous les regroupements d'actes, nous avons pu déterminer la prime pure pour une entreprise spécifique.

Une analyse menée sur 30 entreprises sélectionnées aléatoirement a permis de comparer la prime pure a posteriori calculée avec la méthode bayésienne à la prime pure reconstituée à partir de la cotisation collective, offrant ainsi une évaluation de la pertinence des tarifs réels des entreprises par rapport à leurs niveaux de risque estimés avec les données historiques.

Bien que l'approche bayésienne offre des avantages notables pour la tarification a posteriori, tels que l'intégration de connaissances a priori et une interprétation aisée des modèles et résultats, certains aspects peuvent encore être améliorés. L'utilisation systématique de modèles hiérarchiques bayésiens pour les données démographiques et géographiques pourrait affiner davantage les modèles, tandis que l'application de processus gaussiens pourrait mieux capturer les variations temporelles et les interactions entre les différents actes de soins.

En conclusion, l'approche bayésienne se révèle être un outil pertinent et efficace pour la tarification a posteriori en assurance santé collective, offrant une méthode innovante pour naviguer dans la complexité de ce domaine.

Synthesis note

This thesis delves into the methods of pricing in the field of group health insurance, with a special focus on the bayesian approach for a posteriori pricing. Relying on a substantial volume of real historical data, this study demonstrates, through concrete case studies, the effectiveness and relevance of using bayesian models in a posteriori pricing for group health insurance. It provides a detailed analysis of how these models can be utilized to address the specific challenges of this sector, thereby offering innovative perspectives and advanced solutions for the inherent complexity of actuarial pricing.

Research problem and motivation :

The group health insurance sector, characterized by its constant dynamics, faces significant regulatory challenges and increased competition following the introduction of mid-year cancellation policies. In this competitive environment, it is crucial for insurers to effectively manage risks and adjust their premiums in a responsive and strategically refined manner.

A posteriori pricing, which represents a dynamic approach to adjusting premiums based on the actual claims experience of insured individuals, proves crucial for maintaining profitability and meeting financial and strategic expectations in insurance product design.

However, traditional methods of a posteriori pricing tend to require data aggregation at a relatively general level. Additionally, modern approaches from Data Science, such as generalized linear models or decision tree-based methods, impose strict data structures and have limitations, especially in the absence of past claims.

The bayesian approach then emerges as a promising solution. Its ability to integrate expertise through a priori distributions before even analyzing the data, combined with its flexibility, interpretability, and effective consideration of the effects of the absence of claims, makes it a strategically advantageous choice.

Historically, the application of the bayesian approach has been limited by its computational complexity. However, thanks to methodological advancements such as Markov Chain Monte Carlo (MCMC) methods and technological progress, including the development of software like PyStan and PyMC, this approach has become more accessible. In this context, we chose to apply the bayesian approach to voluminous data for a posteriori pricing, paving the way for innovative solutions in the field of group health insurance.

Followed methodology : The a posteriori pricing in group health insurance, discussed in this thesis, is based on a methodological approach that structures several key stages, aiming to ensure an accurate and relevant analysis of the data in the context of applying bayesian models. Here are the main components of this methodology:

- **Data preparation and working assumptions** : The methodology of this thesis began with an exhaustive preparation of data, a process deemed essential for ensuring the quality

and relevance of information for modeling purposes. This stage involved the collection, analysis, cleaning, and aggregation of data from various sources. It was crucial to recognize and manage the imperfections of the raw data, particularly abnormal or outlier values, and to aggregate them judiciously to maintain a balance between model generalization and the preservation of essential information.

- **Descriptive data analysis** : The descriptive analysis of the data was a fundamental step, providing a comprehensive overview of the statistical distributions of the data, including their minimum and maximum values. This analysis allowed for the assessment of the general shape and trends of the data and the selection of the most appropriate type of distribution for the prior law in bayesian models, a crucial choice for the accuracy and validity of the predictive models.

- **Application of bayesian models** : Following thorough preparation and analysis of the data, the next step was the application of bayesian models. This approach enabled the exploitation of nuances and hidden patterns in the data for accurate a posteriori pricing. This step represented a bridge between the meticulous data preparation and its practical application in the complex world of actuarial science, demonstrating the power of actuarial analysis supported by advanced modeling methods.

Work performed and interpretation of results :

Initially, our study delved into the context of group health insurance, focusing on recent regulations and reforms in this field. We closely examined reports and overall statistics of the health insurance market to gain a deep understanding of the study context.

Subsequently, our focus shifted to analyzing the data from Malakoff Humanis Group, in order to examine the application of the bayesian approach in a posteriori pricing. For this purpose, it was essential to define a specific study perimeter with complete and high-quality data.

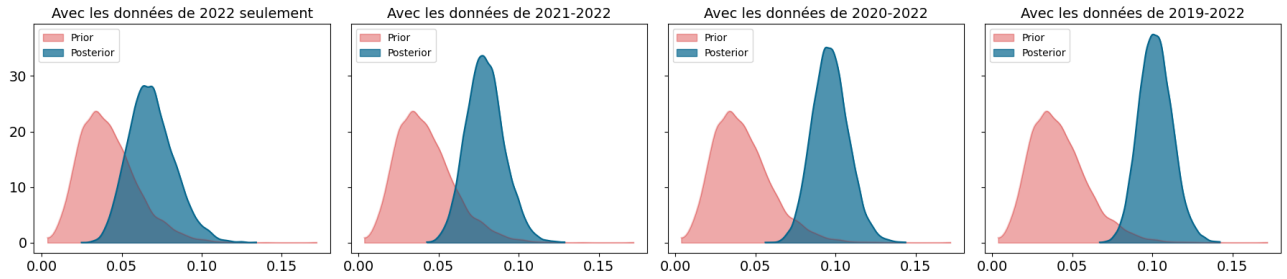
After defining the study scope, we embarked on data collection from various sources and on data analysis. This phase, although time-consuming and demanding, is essential to ensure the quality of the modeling. In this study, more than 10 million lines of data were collected and analyzed.

Particular emphasis was placed on grouping of healthcare acts. Through a analysis of descriptions and criteria such as frequency and amounts, we formed about forty categories of grouping, facilitating descriptive analyses and reducing the data size for a more precise and manageable study.

The data calibration step, based on the proportional presence of beneficiaries, proved crucial for accurate analysis, particularly due to variations in coverage duration within group health insurance contracts.

For each grouping of healthcare acts, two models were developed: one to estimate the annual number of occurrences and another for the associated reimbursement amounts. As depicted in the graph 2 below, the a posteriori distribution, which incorporates historical data, gradually diverges from the a priori distribution, thereby reflecting individual effects.

Figure 2 – Evolution of parameter distribution with data integration



Using the posterior distributions estimated for a grouping of healthcare acts, we simulated the number of healthcare occurrences and the associated reimbursement amounts. The product of these two elements allowed us to estimate the projected costs for the insurer, and thus the pure premium. By summing the posterior pure premiums for all groupings of services, we were able to determine the pure premium for a specific company.

An analysis conducted on 30 randomly selected companies allowed us to compare the pure premium a posteriori calculated using the Bayesian method to the pure premium reconstructed from the actual insurance premium paid, thus providing an assessment of the relevance of the actual company insurance premiums compared to their risk levels estimated from historical data.

Although the bayesian approach offers notable advantages for a posteriori pricing, such as the integration of a priori knowledge and easy interpretation of models and results, certain aspects can still be improved. Systematic use of hierarchical Bayesian models for demographic and geographic data could further refine the models, while the application of gaussian processes could better capture temporal variations and interactions between the different healthcare acts.

In conclusion, the bayesian approach proves to be a relevant and effective tool for a posteriori pricing in group health insurance, offering an innovative method to navigate the complexity of this field.

Remerciements

Avant tout, je tiens à exprimer ma profonde gratitude envers ma manager, Nathalie BRESSON PALEY, responsable actuariat produits collectifs chez Malakoff Humanis, pour la confiance qu'elle m'a accordée et ses précieux conseils. Sa générosité dans le partage de ses connaissances et de son expertise a grandement enrichi mon parcours professionnel et personnel.

Je souhaite également adresser mes sincères remerciements à l'ensemble des membres de l'équipe « Actuariat Produit Standard » pour leur soutien constant, leurs conseils avisés et leur bonne humeur. Un remerciement spécial à Lionel COURTADE et Oussama CHIHI pour leurs conseils techniques avisés et le temps précieux qu'ils m'ont consacré.

Mes remerciements s'étendent aussi aux enseignants de la formation CEA, dont les cours de haute qualité ont été un pilier fondamental de mon apprentissage et de ma compréhension du domaine de l'actuariat.

Enfin, je souhaite exprimer ma gratitude la plus sincère à ma conjointe Chengcheng WEI, pour son soutien indéfectible au quotidien. Sa présence, son encouragement et sa compréhension ont été des sources d'inspiration et de motivation constantes tout au long de ce parcours.

Table des matières

Introduction

Dans un environnement assurantiel en constante évolution, marqué par des réformes réglementaires telles que le contrat responsable, la généralisation de la couverture complémentaire santé, et l'impact significatif de la réforme 100% Santé, l'industrie de l'assurance santé collective se trouve actuellement à une intersection critique, confrontée à des choix stratégiques déterminants pour son avenir. Ces changements, exigeant une réactivité et une précision accrues, soulignent l'urgence d'adapter les méthodes de tarification de manière efficace et innovante. Ce mémoire, en se plongeant au cœur de ces défis et des dynamiques en constante évolution du secteur, vise à examiner en profondeur l'utilisation des modèles bayésiens et leur application concrète dans la tarification a posteriori pour les assurances santé collectives. À travers cette analyse, l'objectif est de mettre en lumière une stratégie novatrice et de proposer des solutions avancées aux assureurs pour naviguer efficacement dans ce contexte complexe et en mutation.

Le premier chapitre de ce mémoire s'immerge dans le contexte assurantiel, en offrant une exploration approfondie du système de santé français et en mettant en lumière l'impact des réformes clés récentes sur le secteur, telles que l'Accord National Interprofessionnel et la réforme 100% Santé. Cette section vise à établir une compréhension solide et détaillée des mécanismes fondamentaux qui régissent les régimes de santé, en soulignant les défis uniques et les complexités inhérentes à leur tarification. Cette analyse initiale pose les bases d'une compréhension complète des enjeux actuels et futurs de l'assurance santé en France.

Dans le second chapitre, l'accent est mis sur la phase cruciale de préparation des données pour la modélisation. Cette partie détaille le périmètre de l'étude, en examinant les types de données utilisées et en explorant les hypothèses fondamentales qui sous-tendent l'analyse. L'étape d'analyse descriptive des données recueillies joue un rôle essentiel, car elle permet de dégager les caractéristiques distinctifs des produits d'assurance santé collective. Cette démarche méthodique et détaillée établit une base solide pour une modélisation précise et approfondie, essentielle pour comprendre et répondre aux besoins spécifiques du secteur de l'assurance santé collective.

Au sein du troisième chapitre, nous abordons l'exploration de l'approche bayésienne appliquée à la tarification a posteriori. Cette section dévoile les principes fondamentaux de cette méthodologie, articulant une synergie entre l'expertise spécialisée du domaine et l'analyse de données pour concevoir des modèles tarifaires dynamiques et adaptés. La portée pratique de ces modèles est illustrée à travers des cas pratiques concrets, qui démontrent non seulement leur efficacité, mais aussi leur pertinence dans le paysage assurantiel actuel. Cette exploration montre comment l'approche bayésienne peut être appliquée pour améliorer les méthodes de tarification a posteriori, en offrant des solutions innovantes adaptées aux défis et aux exigences du secteur.

Le quatrième chapitre s'oriente vers des approfondissements et des perspectives. Il se focalise sur une analyse détaillée des modèles hiérarchiques bayésiens et sur l'examen de l'utilisation des processus Gaussiens dans les modèles bayésiens. Il aborde également divers critères et méthodes pour évaluer l'adéquation des distributions a priori. Cette partie du mémoire vise à approfondir la compréhension, en proposant une analyse plus nuancée et une perspective élargie sur l'application de l'approche bayésienne en matière de tarification a posteriori. Elle met en lumière les aspects susceptibles d'amélioration de

ces modèles avancés pour développer des méthodes de tarification a posteriori encore plus précises et flexibles. De plus, elle explore les implications potentielles de ces améliorations pour l'avenir de la tarification dans le secteur de l'assurance santé collective.

En conclusion, ce mémoire vise à fournir une compréhension approfondie des dynamiques actuelles en assurance santé collective, avec un accent particulier sur l'application de l'approche bayésienne pour la tarification a posteriori. L'objectif est de mettre au point une méthode de tarification non seulement adaptée, mais aussi réactive, pour relever efficacement les défis inhérents à ce domaine. À travers cette étude, nous cherchons à approfondir la compréhension des stratégies de tarification en assurance santé et à offrir des solutions novatrices pour naviguer dans cet environnement complexe.

Chapitre 1

Contexte assurantiel

Ce mémoire explore le domaine du risque santé, un secteur essentiel et en constante évolution. Le premier chapitre se consacre à la mise en contexte de l'environnement assurantiel, en mettant un accent particulier sur l'assurance complémentaire santé. Nous aborderons également les impacts significatifs des contrats responsables ainsi que des réformes de l'Accord National Interprofessionnel *(ANI)* et de la réforme 100% Santé. En outre, nous examinons aussi les répercussions de la pandémie de Covid-19, dont les impacts significatifs sur le secteur de l'assurance santé constituent un élément essentiel à considérer dans le cadre de ce mémoire. Ce chapitre vise à établir une compréhension claire des mécanismes régissant les régimes de santé et des réformes récentes qui ont marqué le secteur, jetant ainsi les bases nécessaires pour une analyse approfondie des dynamiques actuelles et des défis futurs en matière de tarification a posteriori dans ce domaine.

1.1 Description du système de santé français

En France, la couverture de santé s'appuie sur un modèle dual : d'une part, le régime obligatoire, qui est un système mutualisé et étatique géré par la Sécurité sociale, et d'autre part, les régimes complémentaires, qui relèvent du secteur assurantiel privé. Le régime obligatoire fournit une protection santé de base à toute la population, prenant en charge une portion des dépenses médicales. Pour compléter cette couverture initiale, les régimes complémentaires interviennent en couvrant en totalité ou partiellement les frais non pris en charge par le régime obligatoire, réduisant de ce fait les montants restant à la charge des assurés.

Régime obligatoire : la Sécurité sociale

En France, la Sécurité sociale représente un ensemble de dispositifs et d'institutions majoritairement publics visant à protéger les individus contre divers « risques sociaux » tels que la maladie, la maternité/paternité, l'invalidité, le décès, les accidents du travail, les maladies professionnelles, la vieillesse et les besoins familiaux.

Historique : Instituée le 19 octobre 1945 sous le gouvernement de Gaulle, la Sécurité sociale est devenue un pilier du système social public et de l'économie française, principalement financée par les cotisations sociales prélevées sur les salaires, contrairement à des modèles comme le National Health Service britannique, qui repose sur l'impôt. Le régime général de la Sécurité sociale, fondé sur un modèle « bismarckien » et paritaire, est géré conjointement par les partenaires sociaux, notamment les syndicats de travailleurs et les organismes patronaux.

Branches de la Sécurité sociale : La Sécurité sociale se divise en six branches principales[1] : la branche maladie, la branche famille, la branche accidents du travail et maladies professionnelles, la branche retraite, la branche autonomie, et la branche cotisations et recouvrement.

La branche maladie : Cette branche est responsable de la prise en charge des dépenses de santé et garantit l'accès aux soins. Elle gère les risques liés à la maladie, la maternité, l'invalidité, et le

décès, tout en favorisant l'accès à la santé pour les plus démunis et en contribuant au fonctionnement d'établissements médico-sociaux.

En 2022, le régime de base de la Sécurité sociale a versé 221,6 milliards d'euros de prestations nettes dans la branche maladie[2]. Par ailleurs, en 2021, les dépenses totales de santé en France représentaient 12,3% du PIB, soulignant l'importance continue de ce système dans l'économie nationale.

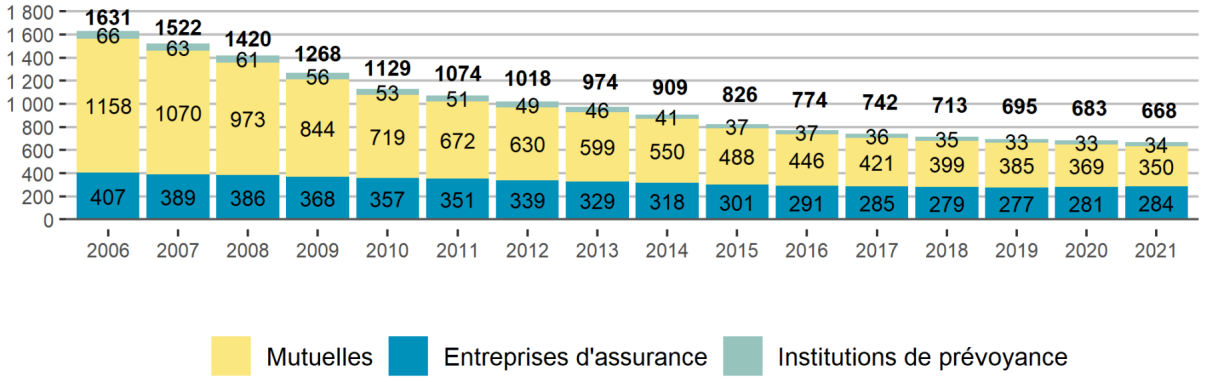
Régimes complémentaires :

Les régimes complémentaires de santé jouent un rôle crucial en complétant les remboursements offerts par la Sécurité sociale. Ils couvrent une partie ou la totalité des dépenses restant à la charge du patient après les remboursements de la Sécurité sociale. Parmi eux, la complémentaire santé solidaire, qui intègre la couverture maladie universelle complémentaire (CMU-C) et l'aide pour une complémentaire santé (ACS), est accessible sous conditions de ressources et vise à garantir une couverture santé plus étendue.

En France, les régimes complémentaires de santé comprennent principalement les sociétés d'assurance, les mutuelles et les institutions de prévoyance, chacune ayant des caractéristiques distinctes qui reflètent des modèles de gestion, de financement et de régulation différents :

- **Sociétés d'assurances** : Ce sont des entreprises de capitaux régies par le code des assurances. Elles se divisent en sociétés d'assurances mutuelles et sociétés anonymes, étant des sociétés commerciales à but lucratif appartenant à des actionnaires.
- **Mutuelles** : Régies par le code de la mutualité, les mutuelles sont définies comme des personnes morales de droit privé à caractère non lucratif. Elles exercent leur activité dans le respect du principe de solidarité et pratiquent une gouvernance démocratique fixée par leurs statuts, prévoyant la participation des membres.
- **Institutions de prévoyance** : Régies par le code de la sécurité sociale et à but non lucratif, les institutions de prévoyance sont établies par accord entre les branches professionnelles et les partenaires sociaux, avec une gestion paritaire entre employeurs et salariés. Elles gèrent principalement la couverture des risques de prévoyance et de santé et certaines cotisations retraite par délégation de la Sécurité sociale.

FIGURE 1.1 – Nombre d'organismes d'assurance agréés par l'ACPR



Le graphique 1.1 illustre clairement la tendance de long terme à la baisse du nombre d'organismes d'assurance en France. En 2021, selon l'Autorité de contrôle prudentiel et de résolution (ACPR)[3], il y avait 668 organismes actifs, dont 350 mutuelles, 284 entreprises d'assurance et 34 institutions de

prévoyance. Cette diminution, observée depuis les années 1990, s'est accélérée après 2013 en raison de nouvelles réglementations européennes et de la mise en place de Solvabilité II. Les fusions et absorptions, encouragées par des exigences réglementaires plus strictes et la généralisation de la complémentaire santé d'entreprise, ont contribué à cette baisse.

Interaction entre régime obligatoire et régimes complémentaires

Le régime obligatoire de santé en France, représenté par la Sécurité sociale, fournit une couverture de base pour divers soins et traitements. Cependant, cette couverture n'est pas exhaustive et peut laisser une partie significative des coûts à la charge des assurés. C'est là que les régimes complémentaires de santé entrent en jeu, offrant des remboursements supplémentaires pour les soins non couverts ou partiellement couverts par le régime obligatoire.

FIGURE 1.2 – Mécanisme de remboursement de frais de santé

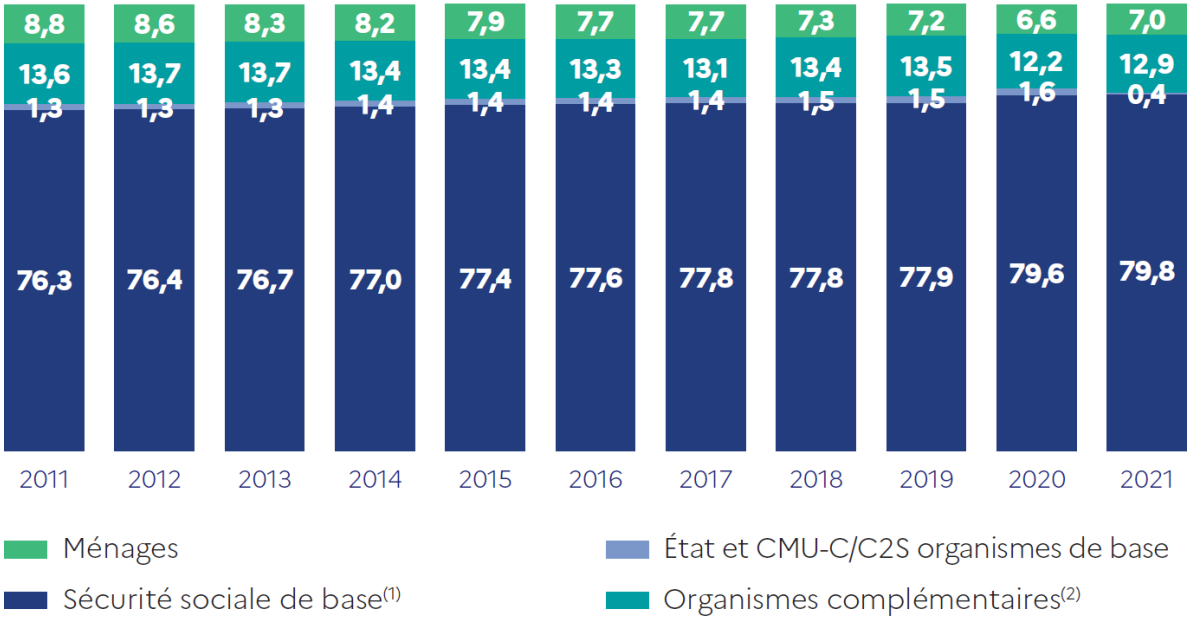
Le schéma 1.2 présenté ci-dessus fournit un aperçu simplifié de la structure de remboursement des frais de santé en France, détaillant spécifiquement la contribution de chaque acteur impliqué. Ce mécanisme, coordonné entre la Sécurité sociale et les complémentaires santé, s'appuie sur plusieurs éléments essentiels :

- **Base de remboursement de la Sécurité sociale *(BRSS)*** : La BRSS est le tarif de référence fixé par la Sécurité sociale pour chaque type de soin. Elle est utilisée pour calculer le montant des remboursements. Par exemple, pour une consultation d'un médecin généraliste, si la BRSS est de 25 euros, c'est sur cette base que le remboursement sera calculé.
- **Ticket modérateur** : Il s'agit de la part des frais médicaux non couverte par la Sécurité sociale et pouvant être prise en charge par les complémentaires santé. Par exemple, si la Sécurité sociale rembourse 70% de la BRSS, le ticket modérateur représente les 30% restants.
- **Dépassement d'honoraires** : Ce sont les frais que certains professionnels de santé peuvent facturer en plus de la BRSS. Ces dépassements ne sont pas remboursés par la Sécurité sociale mais peuvent l'être, en tout ou partie, par les assurances complémentaires santé.
- **Frais réels** : Les frais réels correspondent au montant total des dépenses de santé engagées par l'assuré. Ils incluent la BRSS et les éventuels dépassements d'honoraires.
- **Reste à charge** : Le reste à charge pour l'assuré désigne les frais médicaux non remboursés par la Sécurité sociale et les complémentaires santé. Cela inclut principalement les éventuels dépassements d'honoraires et les services non couverts par ces organismes.

- **Remboursement des complémentaires santé** : Les complémentaires santé remboursent une partie ou la totalité du ticket modérateur et des dépassements d'honoraires, selon le contrat souscrit par l'assuré. Elles peuvent également couvrir d'autres frais non pris en charge par la Sécurité sociale.

Dans la pratique, la coordination entre ces deux types de régimes est cruciale pour assurer une couverture sanitaire complète et efficace. Typiquement, lorsqu'une personne reçoit des soins de santé, le régime obligatoire rembourse d'abord sa part, basée sur la BRSS. Ensuite, les régimes complémentaires interviennent pour rembourser tout ou partie du reste, selon les garanties souscrites par l'assuré. Cette coordination permet d'alléger le fardeau financier des soins de santé pour les individus, en assurant une prise en charge plus large des coûts associés.

FIGURE 1.3 – Évolution de la structure du financement de la consommation de soins et de biens médicaux en pourcentage[2]



(1) Y compris déficit des hôpitaux publics.
(2) Y compris prestations CMU-C/C2S versées par ces organismes.
Source : Comptes de la santé, Drees, 2022.

Le graphique 1.3 illustre la répartition des remboursements des frais de soins de santé en France entre 2011 et 2021, avec la Sécurité sociale, les organismes complémentaires de santé, et la contribution directe des ménages constituant les sources principales de financement. La part de la Sécurité sociale de base a augmenté progressivement de 76,3% en 2011 à 77,9% en 2019, atteignant même 79,6% en 2020 et 79,8% en 2021 en raison de la pandémie de COVID-19. Cela indique que la Sécurité sociale a pris en charge une plus grande proportion des frais de soins de santé au fil du temps. La contribution des ménages a légèrement diminué, passant de 8,8% à 7,2% sur la même période, avec une diminution notable en 2020 à 6,6% avant de remonter à 7,0% en 2021. Il est à noter que la part des organismes complémentaires reste relativement constante à travers les années.

1.2 Les évolutions réglementaires en assurance santé

Les contrats responsables

Les contrats responsables représentent un cadre réglementaire mis en place pour les organismes complémentaires de santé, introduits le 1er janvier 2006, ont été conçus dans le but d'encourager une utilisation plus responsable des soins de santé et de renforcer la couverture des assurés en respectant le « parcours de soins ».

Lorsqu'un assuré suit le parcours de soins, ces contrats couvrent intégralement le ticket modérateur pour la plupart des soins, à l'exception de certains médicaments faiblement remboursés, l'homéopathie et les cures thermales. Ils prennent également en charge le forfait journalier hospitalier sans limitation de durée et prennent en charge, selon les garanties contractuelles, les dépassements d'honoraires pour les médecins adhérents à l'OPTAM *(Option Pratique Tarifaire Maîtrisée)* ou l'OPTAM-CO *(Option Pratique Tarifaire Maîtrisée Chirurgie et Obstétrique)*. Concernant l'optique, ils permettent le remplacement d'un équipement tous les deux ans, avec des conditions particulières pour les mineurs ou en cas de modification de la vue.

Si l'assuré sort du parcours de soins coordonné, les contrats responsables ne remboursent pas les dépassements d'honoraires, ni les majorations du ticket modérateur. Par exemple, lors d'une consultation chez un médecin, le remboursement de la Sécurité sociale est de 70% de la base de remboursement si le patient respecte le parcours de soins coordonnés, mais il est réduit à 30% pour les consultations hors parcours. Toutefois, même dans ce cas de pénalité financière où seulement 30% est remboursé, les contrats responsables ne prennent pas en charge la différence. Ainsi, en cas de non-respect du parcours de soins, le montant restant à la charge de l'assuré peut considérablement augmenter.

Quant aux avantages, les contrats responsables bénéficient d'un cadre fiscal et social privilégié. Pour les employeurs offrant ces contrats de manière collective et obligatoire, les cotisations sont exonérées de charges sociales jusqu'à un certain plafond. Pour tous les contrats santé, la notion de contrat responsable implique aussi un taux réduit de la Taxe de Solidarité Additionnelle *(TSA)* à 13,27% au lieu de 20,27% pour les contrats non responsables. Pour les travailleurs non-salariés, la cotisation est déductible de l'impôt sur le revenu, tandis que pour les salariés, la part salariale des cotisations peut être déduite dans le calcul de l'impôt sur le revenu. Ces incitations fiscales et sociales visent à rendre les contrats responsables plus attractifs pour les employeurs et les travailleurs, favorisant ainsi une meilleure couverture santé et une plus grande responsabilisation dans l'utilisation des services de santé.

L'accord national interprofessionnel

L'accord national interprofessionnel *(ANI)* de 2013, transposé dans la loi n° 2013-504 du 14 juin 2013 relative à la sécurisation de l'emploi[4], a marqué une étape décisive dans la généralisation de la couverture complémentaire santé en France. Cette loi obligeait les employeurs du secteur privé à fournir une couverture santé complémentaire à tous leurs salariés à partir du 1er janvier 2016, avec des garanties au moins aussi avantageuses que celles établies par le code de la sécurité sociale et une prise en charge d'au moins la moitié des cotisations par l'employeur. Les négociations entre les employeurs et les délégués syndicaux devaient définir le contenu des garanties, la répartition des cotisations, et le choix de l'assureur. Cette réforme visait à améliorer l'accès à la santé pour tous les salariés, en leur garantissant une couverture santé d'entreprise dont une partie était prise en charge par l'employeur.

- **Panier de soins** : En l'absence d'accord spécifique, l'ANI a établi un socle de garanties minimales pour les mutuelles d'entreprise. Ce panier de soins minimal doit couvrir le ticket modérateur pour les soins pris en charge par le régime obligatoire, le forfait journalier hospitalier, 125% de la BRSS pour les frais dentaires, et un forfait minimal de 100 € pour les frais d'optique.
- **Participation de l'employeur** : La loi impose aux entreprises de financer au moins 50% du coût du contrat de mutuelle. Cette contribution peut être augmentée en vertu d'une convention collective ou d'un accord de branche.
- **Portabilité des garanties** : La loi a également renforcé la portabilité des garanties de mutuelle d'entreprise, permettant aux salariés de conserver leur complémentaire santé collective après avoir quitté l'entreprise, sous certaines conditions, notamment en cas de rupture de contrat de travail.

La réforme 100% Santé

La réforme « 100% Santé », anciennement connue sous le nom de « reste à charge zéro », a été introduite pour garantir un accès équitable à des soins essentiels dans les domaines de l'optique, du dentaire et de l'audiologie. Cette initiative visait à réduire les disparités d'accès aux soins de santé, face à la difficulté rencontrée par certains citoyens français à se procurer des équipements médicaux onéreux, tels que les lunettes de vue, les prothèses dentaires et les aides auditives. En effet, parmi les 20% de Français les plus modestes, près d'un sur cinq renonçait à s'équiper en optique, et près d'un sur trois à réaliser des soins dentaires, soulignant ainsi un enjeu à la fois sanitaire et social.

Cette réforme, qui s'est déployée progressivement sur trois ans, repose sur un système de prise en charge intégrale de certains soins par la Sécurité sociale et les complémentaires santé. Cela signifie que pour les soins inclus dans le « panier » de la réforme, les patients n'auront rien à payer de leur poche. Pour y parvenir, les tarifs de ces soins ont été plafonnés, et les bases de remboursement de la Sécurité sociale ont été progressivement augmentées. De plus, les garanties des contrats de santé responsables des complémentaires ont été adaptées pour éliminer tout reste à charge pour ces prestations. Les offres du panier « 100% Santé »[5] sont composées comme suit :

- **Optique** :
  - Montures respectant les normes européennes, avec un prix inférieur ou égal à 30 €, disponibles en au moins 17 modèles différents et en deux coloris.
  - Verres capables de traiter tous les troubles visuels, avec amincissement adapté au trouble, durcissement pour éviter les rayures, et traitement anti-reflet obligatoire.
- **Audiologie** :
  - Tous les types d'appareils auditifs, y compris les contours d'oreille classiques, les contours à écouteur déporté, et les intra-auriculaires.
  - Au moins 12 canaux de réglage (ou dispositif de qualité équivalente) pour une correction adaptée au trouble auditif, et système d'amplification des sons extérieurs restitué à au moins 30 dB.
  - Une garantie de 4 ans et au moins trois options parmi un système anti-acouphène, connectivité sans fil, réducteur de bruit du vent, synchronisation binaurale, directivité microphonique adaptative, bande passante élargie $\geq$ 6000 Hz, fonction d'apprentissage de sonie, et

système anti-réverbération.

- **Dentaire** :
  - Couronnes céramiques monolithiques et céramo-métalliques pour les dents visibles, et des couronnes métalliques pour toutes les dents.
  - Inlays-core et couronnes transitoires.
  - Bridges céramo-métalliques pour le remplacement d'une incisive, bridges entièrement métalliques pour toutes les dents, et prothèses amovibles à base de résine.

La résiliation infra-annuelle

La résiliation infra-annuelle *(RIA)* s'applique tant aux contrats individuels qu'aux contrats collectifs de complémentaire santé. Elle était initialement non autorisée, puisqu'une couverture complémentaire santé est généralement souscrite pour une durée minimale de douze mois et renouvelée par tacite reconduction. Toutefois, la loi n° 2019-733 du 14 juillet 2019, entrée en vigueur le 1er décembre 2020, a introduit une modification significative, permettant aux assurés de résilier leur contrat après un an de souscription, à tout moment, sans frais ni pénalité.

Le décret n° 2020-1438 du 24 novembre 2020 relatif au droit de résiliation sans frais de contrats de complémentaire santé, établit les modalités d'application de la loi du 14 juillet 2019. Ce décret spécifie les types de contrats éligibles à cette faculté. Il s'agit des contrats qui couvrent les risques liés à la santé, notamment en remboursant les frais occasionnées par des situations telles que la maladie, la maternité ou les accidents.

Depuis le 20 mars 2022, la résiliation infra-annuelle s'applique également aux contrats de complémentaire santé qui couvrent le risque de perte d'autonomie. Cette extension a été mise en place suite à un décret du 17 mars 2022.

1.3 Nouveaux défis en assurance santé : RGPD et Covid-19

Le règlement général sur la protection des données

Le règlement général sur la protection des données *(RGPD)*, mis en place par l'Union européenne depuis le 25 mai 2018, représente une étape majeure dans la réglementation de la protection des données personnelles. Cette législation vise à renforcer les droits des individus sur leurs données personnelles et à harmoniser les règles de protection des données au sein de l'UE. Voici les points clés du RGPD :

- **Consentement explicite** : Les organisations doivent obtenir un consentement clair et explicite des personnes avant de collecter ou de traiter leurs données personnelles.
- **Droit à l'oubli** : Les individus peuvent demander la suppression de leurs données personnelles dans certaines circonstances.
- **Transfert de données** : Le RGPD restreint le transfert de données personnelles vers des pays hors de l'UE, sauf si ces pays assurent un niveau de protection adéquat.
- **Délégué à la protection des données *(DPO)*** : Certaines organisations doivent désigner

un DPO pour superviser la conformité au RGPD.

- **Droits des sujets de données** : Les individus ont divers droits, comme accéder à leurs données, les corriger, les transférer, ou s'opposer à leur traitement.
- **Sanctions** : Le non-respect du RGPD peut entraîner des amendes importantes, pouvant atteindre 4% du chiffre d'affaires annuel mondial de l'organisation ou 20 millions d'euros, selon le montant le plus élevé.

Dans le secteur de l'assurance, qui manipule une multitude de données personnelles, y compris des informations sensibles liées à la santé, le RGPD a imposé des exigences accrues. Les assureurs, en tant que responsables du traitement des données, sont tenus de respecter les principes de protection des données, de transparence et de responsabilité. Cela inclut le consentement éclairé des assurés pour le traitement de leurs données sensibles, une utilisation limitée à des fins spécifiques et clairement définies, ainsi que la mise en œuvre de mesures de sécurité robustes pour prévenir les accès non autorisés ou les fuites de données.

La réglementation impose également une transparence accrue vis-à-vis des assurés. Les assureurs doivent non seulement informer clairement les clients sur l'utilisation de leurs données, mais aussi respecter leurs droits en matière d'accès, de rectification, d'opposition, d'effacement et de portabilité des données. Ces droits renforcent le contrôle des individus sur leurs informations personnelles, donnant par exemple la possibilité de récupérer leurs données dans un format utilisable ou de les transférer à une autre compagnie d'assurance.

Un aspect crucial du RGPD est la limitation de la durée de conservation des données. Les données ne peuvent être conservées que pour une période nécessaire et justifiable, souvent alignée sur les délais de prescription en justice, généralement fixés à 5 ans. Cette limitation présente un défi pour les assureurs, en particulier pour la modélisation des risques. L'accès restreint à des données historiques longues peut introduire un biais dans les modèles actuariels, affectant potentiellement la précision des évaluations de risque et des tarifications.

La pandémie de Covid-19

En France, la pandémie de Covid-19 a entraîné des répercussions majeures, en particulier en ce qui concerne les confinements et l'impact sur les assureurs santé complémentaires.

- **Les confinements en France** :

Les confinements en France ont été mis en place par le gouvernement français à trois reprises pour freiner la propagation du virus. Les périodes de confinement étaient les suivantes :

- Du 17 mars au 11 mai 2020 (1 mois et 25 jours)
- Du 30 octobre au 15 décembre 2020 (1 mois et 15 jours)
- Du 3 avril au 3 mai 2021 (28 jours)

Ces confinements impliquaient des restrictions sévères sur les déplacements, limités au strict nécessaire comme les courses alimentaires, les soins médicaux, le travail en cas d'impossibilité de télétravail, les sorties à proximité du domicile pour des activités physiques individuelles ou pour la garde d'enfants. Sur le plan économique, les confinements ont entraîné la fermeture temporaire des magasins et des entreprises considérés comme "non essentiels", ainsi que des lieux de sociabilité tels que les bars, restaurants, et autres établissements de loisirs.

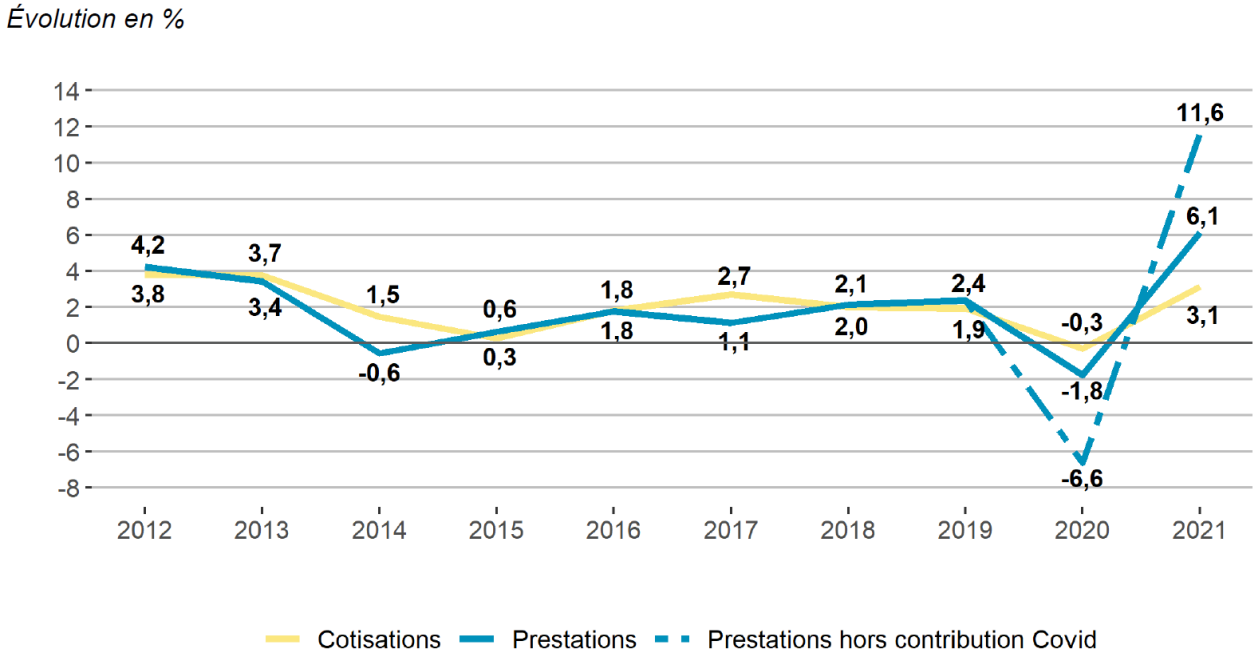
- **Taxe Covid pour les assureurs santé complémentaires** :

En réponse aux défis économiques posés par la pandémie, le gouvernement français a imposé une taxe exceptionnelle sur les complémentaires santé. Cette taxe, d'un montant de 1,5 milliard d'euros, a été instaurée pour compenser les pertes massives de la Sécurité sociale, qui a dû faire face à des dépenses importantes liées à la crise du Covid-19.

Durant les périodes de confinement, les mutuelles de santé ont réalisé d'importantes économies, estimées à plus de 2,6 milliards d'euros, en raison du renoncement aux soins et de la prise en charge à 100% par la Sécurité sociale de certains actes comme les tests de dépistage et les téléconsultations. Toutefois, ces économies pourraient être réévaluées à la baisse en raison du rattrapage des soins après les périodes de confinement. La taxe exceptionnelle sur les complémentaires santé, bien que nécessaire selon le gouvernement pour rétablir un équilibre financier, a été source de controverse au sein du secteur assurantiel. La taxe exceptionnelle a été discutée en juin 2020, actée en août, et devait être réglée sur deux ans, en 2021 et en 2022.

- **Fort rebond des prestations 2021** :

FIGURE 1.4 – Évolution des cotisations et prestations en santé



En 2021, les prestations des organismes complémentaires en France ont connu une augmentation marquée ($+6,1\%$)[3], stimulée par l'effet de rattrapage suite à la baisse significative de 2020 ($-1,8\%$) et renforcée par la réforme 100% Santé, particulièrement avec l'augmentation des remboursements dentaires. Cette hausse est plus importante que celle des cotisations ($+3,1\%$), qui n'a pas compensé la forte augmentation des charges de prestations. Sans tenir compte de la contribution exceptionnelle liée à la COVID-19, les prestations auraient chuté plus nettement en 2020 ($-6,6\%$) et auraient augmenté plus fortement en 2021 ($+11,6\%$). Sur deux ans, de 2019 à 2021, l'augmentation globale des prestations est de $+4,2\%$, reflétant une dynamique de croissance malgré la crise sanitaire.

Le graphique 1.4 illustre ces tendances, montrant une baisse des prestations en 2020 suivie d'une forte reprise en 2021. On observe également que les mesures de gestion des dépenses prises par les organismes complémentaires entre 2014 et 2017, comme la modération des dépenses en optique, ont ralenti la croissance des prestations par rapport à 2012-2013.

Conclusion

Dans ce premier chapitre "Contexte assurantiel", nous avons exploré les divers aspects de l'assurance santé en France, un secteur caractérisé par son évolution constante et la complexité de ses interactions entre le régime obligatoire et les régimes complémentaires. Les contrats responsables et l'Accord National Interprofessionnel représentent des jalons importants dans la régulation de ce domaine, tandis que la réforme 100% Santé souligne son orientation vers une accessibilité accrue aux soins.

Par ailleurs, l'avènement de la résiliation infra-annuelle a engendré une dynamique de concurrence plus marquée. Cependant, l'émergence de la pandémie de Covid-19, conjuguée aux réformes récentes, a créé un climat d'incertitude pour les assureurs santé, les confrontant à des difficultés dans la maîtrise de leurs portefeuilles. En outre, la mise en application du RGPD impose aux assureurs de nouvelles exigences quant à la collecte, au traitement et à la conservation des données personnelles, accentuant la complexité de la gestion des risques dans un environnement déjà en mutation.

Cette période de changement et d'instabilité met en lumière la nécessité d'une tarification réfléchie et pertinente, capable de répondre aux défis de la gestion des risques et de la concurrence dans le marché, dans un contexte où les enjeux économiques, sanitaires et sociaux sont étroitement entrelacés. C'est dans cette toile de fond complexe que s'ancre l'objet de ce mémoire, portant sur l'étude approfondie de la tarification a posteriori en assurance santé collective.

Chapitre 2

Préparation des données en vue de la modélisation

2.1 Le périmètre de l'étude

Malakoff Humanis, un acteur majeur dans le domaine de la protection sociale en France, a été formé en janvier 2019 suite au rapprochement des groupes Malakoff Médéric et Humanis. Avec 8,2 milliards d'euros de fonds propres[6] en 2022, le groupe s'est établi comme un pilier dans ce secteur. En assurance, il sert plus de 371 000 entreprises clientes et protège environ 10 millions de personnes, incluant les assurés et leurs ayants droit [1].

La structure de Malakoff Humanis, à l'instar d'autres groupes assurantiels, est caractérisée par sa complexité, reflétant la diversité et l'étendue de ses activités. Cette complexité se manifeste dans l'organisation de ses différentes entités et dans la variété des services qu'elles proposent. Le schéma 2.1 détaillé ci-dessous offre une représentation visuelle de cette structure, permettant une meilleure compréhension de ses différentes composantes et de leur interconnexion.

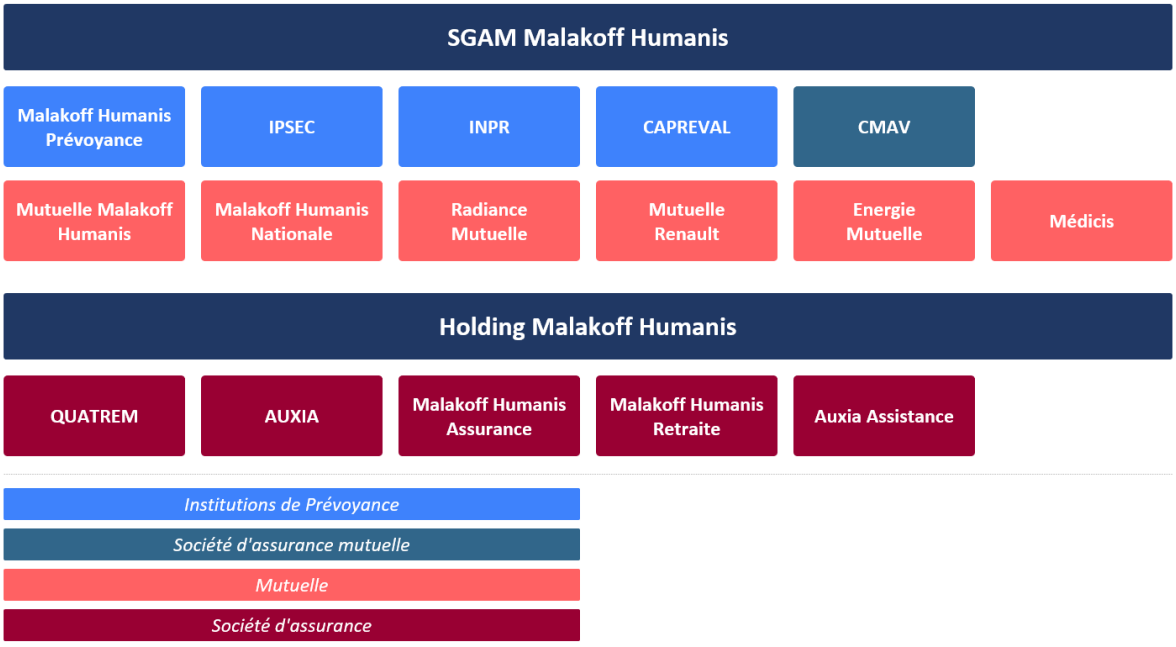


FIGURE 2.1 – Structure du groupe Malakoff Humanis dans le secteur assurantiel

Cette étude se focalise sur l'analyse des méthodes de tarification a posteriori dans le cadre des contrats de complémentaire santé collective. Afin de cerner avec précision le périmètre de notre recherche, nous nous penchons sur l'assureur « Malakoff Humanis Prévoyance », une entité juridique

1. Malakoff Humanis - Chiffres clés au 31/12/2022

distincte au sein du groupe Malakoff Humanis. Cette sélection n'est pas arbitraire mais répond à un besoin de simplification et d'efficacité dans la gestion des données. En se concentrant sur Malakoff Humanis Prévoyance, qui gère un nombre important de contrats santé collectifs en gestion directe, nous disposons d'un volume important de données, offrant une base solide pour des analyses détaillées et une modélisation robuste.

De plus, au sein de cette entité, notre attention se porte spécifiquement sur un produit de santé collective interprofessionnel. Cette focalisation sur un produit spécifique nous permet de limiter les difficultés liées à la qualité des données et de réduire la complexité inhérente à l'étude de plusieurs produits. En concentrant notre analyse sur ce produit particulier, nous sommes en mesure de mieux saisir et d'examiner en profondeur les processus de tarification a posteriori propres à ce segment. Cette démarche ciblée est cruciale pour élaborer des modèles bayésiens qui soient à la fois précis et adaptés aux spécificités de ce marché. Une telle approche nous permet d'assurer que les modèles développés sont non seulement théoriquement solides, mais aussi pratiquement applicables et pertinents pour le domaine de la complémentaire santé collective.

Produit : PEPS ÉCO Active Responsable

Le produit santé complémentaire collective « PEPS ÉCO Active Responsable » *(EAR)* proposé par Malakoff Humanis représente une solution pragmatique et complète, conçue pour s'aligner avec les exigences des nouvelles réglementations comme l'ANI de 2016 et le cadre du contrat responsable. Cette offre s'inscrit dans une démarche de confiance et de conformité, assurant une protection santé adaptée aux réalités actuelles des entreprises et de leurs salariés.

Avec PEPS EAR, les entreprises bénéficient d'un accompagnement personnalisé qui leur permet de trouver l'équilibre idéal entre les garanties collectives obligatoires et les options individuelles facultatives. Ce produit tient compte de la diversité régionale en matière de coûts de soins, permettant ainsi une tarification ajustée et pertinente. De plus, il offre la possibilité d'ajouter une surcomplémentaire, donnant aux salariés la liberté de personnaliser davantage leur couverture sans engendrer de charges supplémentaires pour l'employeur.

Du côté des salariés, ce produit représente une opportunité d'accéder à des conditions tarifaires favorables propres aux contrats collectifs, et de bénéficier d'une extension de leur couverture à leur famille. Par ailleurs, l'accès à une surcomplémentaire individuelle leur permet d'affiner leur protection et de mieux répondre à leurs besoins de santé personnels.

Avant de plonger dans l'analyse des données chiffrées, il est essentiel de comprendre la structure du produit PEPS EAR. Ce produit se décline en trois niveaux de couverture distincts pour s'adapter à différents besoins et budgets. Le niveau « Essentielles » représente l'offre de base, proposant des garanties fondamentales à un coût accessible. Le niveau intermédiaire, « Maîtrisées », élève le degré de protection avec un équilibre entre étendue des garanties et modération des coûts. Enfin, « Renforcées », le niveau supérieur, offre une couverture exhaustive pour ceux qui recherchent une protection maximale. Chaque niveau comprend plusieurs formules, qui correspondent à des ensembles de garanties spécifiques. Une fois une formule sélectionnée, tous les assurés bénéficient du même niveau de garanties, garantissant ainsi une uniformité au sein de chaque option. Les graphiques 2.2 et 2.3 illustrent l'évolution de 2019 à 2022, des cotisations et du nombre d'entreprises ayant opté pour ces différentes formules, offrant un aperçu précis de la dynamique du produit et de son accueil sur le marché [1].

1. Les données présentées se concentrent uniquement sur les affiliations obligatoires afin de simplifier la visualisation.

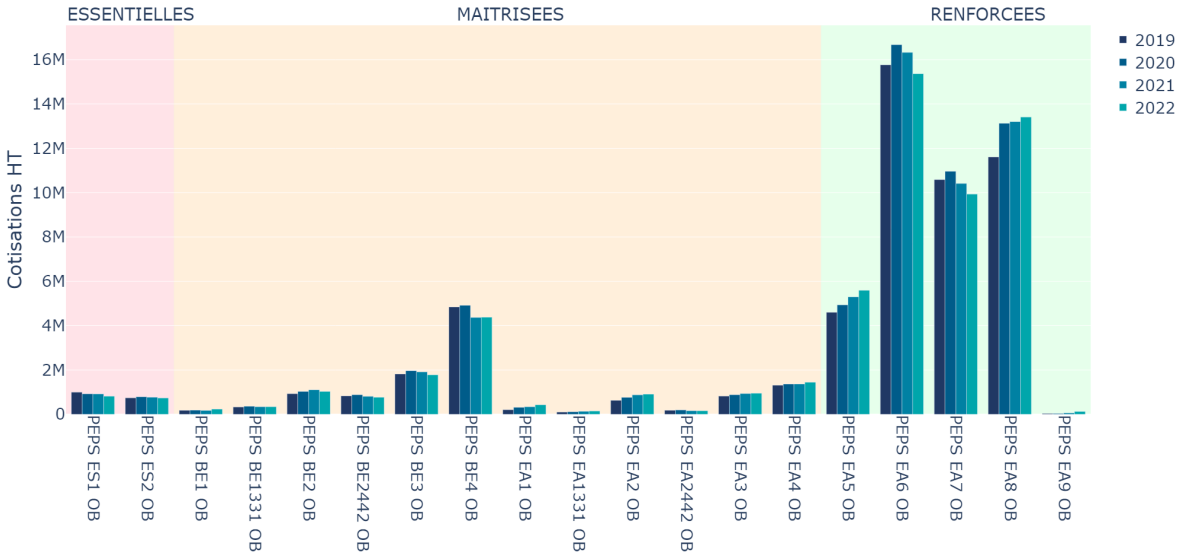


FIGURE 2.2 – Somme des cotisations par formule des produits PEPS EAR 2019-2022

Les données extraites des graphiques 2.2 et 2.3 sur la période de quatre ans montrent une stabilité remarquable en termes de cotisations et du nombre d'entreprises souscrites aux différentes formules du produit PEPS EAR. Les formules appartenant à la gamme ESSENTIELLES, bien que tarifées plus bas et donc générant moins de cotisations, maintiennent un attrait significatif en termes de nombre d'entreprises couvertes. Cette performance est conforme aux attentes pour une entrée de gamme, qui vise à offrir une couverture essentielle à un coût réduit. En revanche, la formule « PEPS EA6 OB » affiche une performance différente, dominant le tableau avec les cotisations les plus élevées et le plus grand nombre d'entreprises couvertes sur les quatre années observées. Ce constat indique une préférence marquée pour cette formule, qui pourrait s'expliquer par un juste équilibre entre l'étendue de la couverture offerte et les coûts associés, la rendant ainsi particulièrement attrayante pour une large palette d'entreprises.

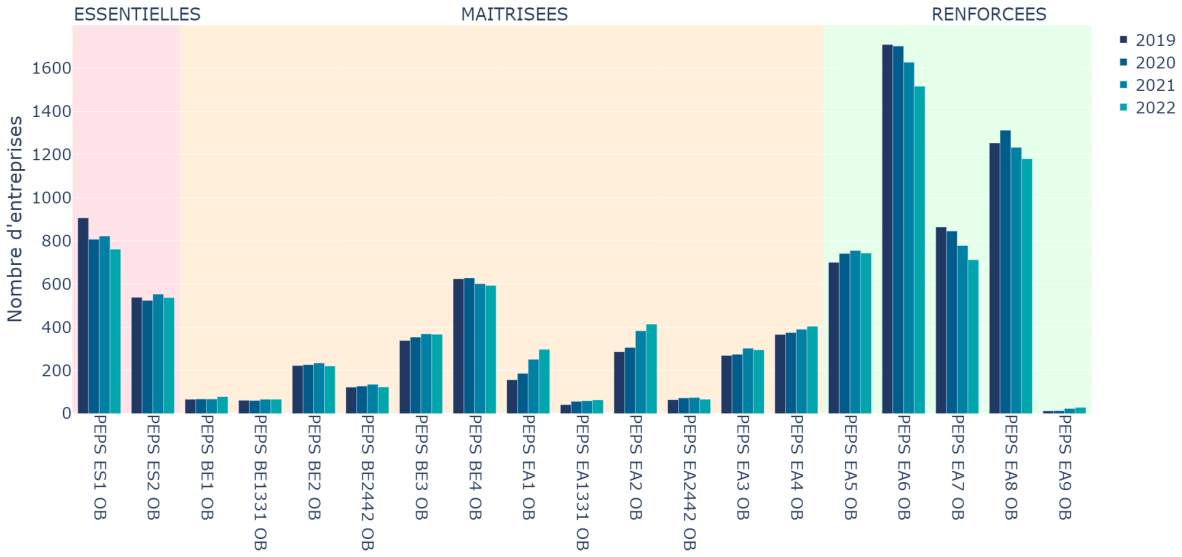


FIGURE 2.3 – Répartition des entreprises par formule des produits PEPS EAR 2019-2022

Dans la poursuite de ce mémoire, notre attention se portera sur la formule « PEPS EA6 OB » en tant que cas d'étude principal. Cette formule se distingue par sa popularité élevée et son haut niveau de couverture, offrant ainsi un volume de données significatif, idéal pour procéder à une analyse bayésienne approfondie.

2.2 La base de données et des hypothèses

Dans le cadre de l'étude sur la tarification a posteriori pour les contrats santé collectifs, nous nous appuyons sur une base de données interne riche et structurée. Cette base de données est essentielle pour comprendre et modéliser les risques associés aux produits de santé collective.

Présentation des tables de données utilisées

Parmi les nombreuses tables disponibles dans notre base de données, nous avons sélectionné quatre tables clés pour nos analyses et modélisations. Ces tables indispensables sont : la table des contrats collectifs, la table des cotisations collectives, la table des bénéficiaires collectifs, et la table des prestations détaillées.

Table des contrats collectifs

La table des contrats collectifs est un élément fondamental de notre base de données. Elle recense chaque contrat de santé collective souscrit par les entreprises. Pour chaque contrat, les informations enregistrées incluent :

- **Organisme assureur** : Désigne la compagnie d'assurance ou l'organisme qui fournit la couverture d'assurance.
- **Produit** : Réfère au produit d'assurance proposé par l'organisme assureur.
- **Type de garantie** : Spécifie le type de couverture offerte par le contrat, par exemple santé ou prévoyance.
- **Formule** : Détaille le niveau de couverture offert par le produit.
- **Numéro d'entreprise** : Numéro unique identifiant l'entreprise souscriptrice du contrat.
- **Code NAF** : Code d'activité française qui classifie l'activité principale exercée par l'entreprise.
- **Numéro de contrat** : Identifiant unique attribué à chaque contrat d'assurance.
- **Date d'effet de contrat** : La date à laquelle le contrat a pris effet.
- **Date de radiation** : La date à laquelle le contrat prend fin, soit par changement de contrat, résiliation, ou autre motif.
- **Famille population** : Catégorise les bénéficiaires selon leur statut professionnel (cadre, non-cadre) ou regroupe l'ensemble du personnel.

- **Mode d'affiliation** : Indique si l'adhésion au contrat est obligatoire ou facultative pour les bénéficiaires.

Grâce à la table des contrats collectifs, nous avons la capacité de sélectionner de manière précise et ciblée tous les contrats pertinents pour notre étude. En particulier, nous concentrons notre attention sur les contrats dont l'organisme assureur est Malakoff Humanis Prévoyance, avec une attention spécifique portée au produit PEPS EAR et à la formule EA6. Cette sélection nous permet d'affiner notre périmètre d'étude sur un segment spécifique et pertinent de notre portefeuille.

De plus, pour chaque entreprise concernée, en examinant les dates d'effet de contrat et de radiation, nous sommes en mesure de calculer le taux de présence de ces entreprises sur différentes années. Cette approche nous aide à éliminer les entreprises qui ont résilié leurs contrats depuis longtemps ainsi que les nouvelles souscriptions qui manquent de données historiques. Ces critères de sélection nous assurent de travailler sur un ensemble de données cohérent et représentatif, où chaque contrat inclus dans l'étude apporte une valeur significative et pertinente à notre modélisation actuarielle.

Table des cotisations collectifs

La table des cotisations collectifs fournit des informations détaillées sur les cotisations perçues pour les contrats de complémentaire santé. Voici les variables clés de cette table :

- **Numéro de contrat** : Identifiant unique du contrat permettant de lier les données entre les différentes sources.
- **Type de cotisation** : Catégorise le type de cotisation selon le choix de l'entreprise et la structure familiale l'assuré. La structure de cette variable se présente sous diverses formes :
  - **Tarif unique famille** : Un tarif unique est appliqué à l'ensemble des salariés, indépendamment de la situation familiale de l'assuré (célibataire, marié, avec ou sans enfant).
  - **Adulte/Enfant** : La cotisation est ajustée en fonction de la composition familiale, avec des tarifs distincts pour les adultes et les enfants.
  - **Isolé/Famille** : Offre une tarification différenciée selon que l'assuré est sans conjoint et sans enfant (tarif « isolé ») ou s'il a une famille (tarif « famille »).
- **Survenance** : Indique l'année à laquelle la cotisation est devenue due.
- **Début période facturée** : Marque le début de la période pour laquelle la cotisation est facturée.
- **Fin période facturée** : Indique la fin de la période de couverture pour laquelle la cotisation est perçue.
- **Montant** : Le montant total de la cotisation facturée pour la période spécifiée.

Cette table des cotisations collectives renseigne sur les montants des cotisations payées par les entreprises, reflétant le niveau de tarification a priori établi lors de la souscription. Les tarifs initiaux sont déterminés en prenant en compte des variables globales propres à chaque entreprise, telles que l'âge moyen des employés, la zone géographique de l'entreprise et d'autres facteurs pertinents.

L'intérêt majeur de cette table réside dans sa capacité à reconstituer la prime pure des entreprises,

ce qui nous permet de la comparer avec la prime pure calculée à partir des modèles bayésiens. La comparaison est essentielle pour évaluer la pertinence de la tarification initiale. En analysant les écarts entre ces deux types de tarification, nous pouvons juger de l'efficacité et de l'adéquation du modèle de tarification appliqué, et ainsi apporter les ajustements nécessaires pour une tarification plus juste et plus adaptée.

Table des bénéficiaires collectifs

Cette table offre des informations détaillées concernant les assurés et leurs ayants droit. Elle contient des données clés telles que :

- **Numéro de contrat** : Identifiant unique du contrat permettant de lier les données entre les différentes sources.
- **Numéro d'assuré** : Identifiant unique attribué à chaque assuré principal sous le contrat.
- **Date de naissance de l'assuré** : Indique la date de naissance de l'assuré principal.
- **Numéro de bénéficiaire** : Identifiant unique pour chaque bénéficiaire couvert par le contrat, y compris l'assuré principal et ses ayants droit.
- **Date de naissance du bénéficiaire** : Indique la date de naissance du bénéficiaire.
- **Lien du bénéficiaire** : Catégorise le lien familial entre l'assuré principal et le bénéficiaire (conjoint ou enfant).
- **Date d'affiliation du bénéficiaire** : Date à laquelle chaque bénéficiaire commence à être couvert par le contrat.
- **Date de sortie du bénéficiaire** : Date à laquelle un bénéficiaire cesse d'être couvert par le contrat.

La table des bénéficiaires collectifs est un élément essentiel pour cette étude, car elle fournit une vue d'ensemble détaillée de la démographie des bénéficiaires couverts par les contrats de santé collective. Cette table contient des informations précises telles que l'âge des bénéficiaires et leurs périodes d'affiliation, permettant une analyse approfondie de la population assurée. Pour chaque bénéficiaire, il est possible de calculer la durée de présence au prorata pour chaque année. Cette mesure est particulièrement importante car elle permet de déterminer les fréquences annuelles des actes de soins, un indicateur clé pour l'évaluation du risque et l'analyse des coûts.

Table des prestations détaillées

La table des prestations est la plus importante pour cette étude, car elle contient des informations détaillées sur les actes de soins de santé réalisés par les bénéficiaires. Cette table est cruciale pour analyser les coûts des soins, les comportements de consommation médicale, et pour tout calcul actuariel et la modélisation des risques. Les variables principales de cette table sont les suivantes :

- **Numéro de contrat** : Identifiant unique du contrat permettant de lier les données entre les différentes sources.
- **Numéro de bénéficiaire** : Identifiant unique du bénéficiaire ayant reçu les soins.

- **Date d'acte de soins** : Date à laquelle l'acte médical a été réalisé.
- **Libellé d'acte de soins** : Description de l'acte médical réalisé. Cette information est essentielle pour comprendre la nature des soins et pour des analyses spécifiques par type de soin.
- **Dépense réelle** : Montant total dépensé pour l'acte de soins.
- **Base de remboursement** : Tarif de référence fixé par la Sécurité sociale pour chaque type de soin.
- **Taux de remboursement par la Sécurité sociale** : Pourcentage de la base de remboursement pris en charge par le régime obligatoire.
- **Montant de remboursement régime obligatoire** : Montant effectivement remboursé par la Sécurité sociale pour un acte de soins.
- **Montant de remboursement régime complémentaire** : Montant remboursé par la complémentaire santé, qui couvre la part non prise en charge par la Sécurité sociale.
- **Montant de remboursement autre mutuelle** : Montant remboursé par toute autre mutuelle impliquée, le cas échéant.

Cette table des prestations est fondamentale pour nos analyses et notre modélisation. En fournissant des informations détaillées sur chaque acte de soins, elle nous permet d'analyser en profondeur l'évolution des fréquences et des coûts des soins de santé. C'est une source de données précieuse pour examiner les tendances de consommation des soins de santé au sein de la population couverte.

L'analyse des données de cette table nous aide à comprendre non seulement comment les coûts de santé et les comportements évoluent, mais aussi à identifier les différences entre les diverses entreprises qui présentent des profils de risques et des comportements de consommation hétérogènes. Cette compréhension approfondie est essentielle pour affiner notre modèle de tarification a posteriori. Elle nous permet d'ajuster notre modèle en fonction des réalités observées, assurant ainsi que notre tarification reste juste, équilibrée et adaptée aux besoins et aux risques de la population assurée.

Préparation des données et des hypothèses

Dans le cadre de cette étude, la préparation des données est une étape cruciale qui implique la collecte, l'analyse, le nettoyage et l'agrégation des données issues de diverses tables. Cette démarche est indispensable pour garantir la qualité et la pertinence des données en vue de l'étape de modélisation. Il est important de reconnaître que les données brutes réelles sont rarement parfaites. Elles comportent souvent des valeurs anormales ou aberrantes. Dans la plupart des cas, ces anomalies représentent des occurrences rares et peuvent donc être éliminées pour préserver l'intégrité des analyses. De plus, il est parfois nécessaire de limiter les valeurs minimales et maximales des variables pour éviter des distorsions statistiques.

En outre, compte tenu de la richesse et de la précision des informations recueillies, il est essentiel de les agréger judicieusement. Un équilibre doit être trouvé pour éviter la complexité excessive des modèles, qui pourrait empêcher leur généralisation, tout en s'assurant de ne pas perdre d'informations essentielles par une agrégation trop importante. Nous présentons ensuite les hypothèses spécifiques adoptées pour la préparation des données dans le cadre de notre étude.

Regroupement des actes de soins

Dans la table des prestations, comme mentionné précédemment, nous disposons d'une quantité importante d'informations détaillées, la variable « Libellé d'acte de soins » présentant plus de 200 modalités différentes. De plus, en raison des changements réglementaires et des ajustements de gestion interne, ces libellés peuvent varier d'une année à l'autre. Afin d'effectuer une analyse cohérente des données historiques et d'évaluer de manière précise les coûts ainsi que les comportements liés aux soins, il est nécessaire de regrouper les actes de soins. Ce regroupement nous permet d'obtenir une base de données moins complexe pour l'analyse et la modélisation, tout en assurant la cohérence des informations d'une année sur l'autre.

La réalisation de ce regroupement s'est basée sur une analyse des libellés d'actes de soins dans notre base, en tenant compte de critères tels que le nombre de comptages et les montants moyens, médians, minimums et maximums, et surtout en se concentrant sur la nature des actes. Sous l'hypothèse que les actes de soins au sein d'un même regroupement présentent une nature similaire ou que leurs fréquences ou coûts sont comparables, nous avons formé une quarantaine de catégories de regroupement. Par exemple, le tableau 2.1 ci-après illustre le regroupement pour les « Consultations généralistes ». Ce processus de regroupement facilite grandement les analyses descriptives, tout en réduisant de façon significative la taille des données finales, ce qui rend l'étude plus maniable et précise.

Table 2.1 – Exemple de regroupement des actes de soins - Consultations Généralistes

Regroupement des actes	**Libellé d'acte de soins**
Consultation Généraliste	Consultation Généraliste
	Visite Généraliste
	Majo coordination généraliste
	Majoration jour férié généraliste
	Majoration nuit généraliste
	Majoration Jour Férié Généraliste
	Majoration nuit Généraliste
	Majoration minuit six heures
	Déplacement Généraliste
	Dép Généraliste crit. médicaux
	Dép géné nuit crit médicaux
	Dép géné ferié crit médicaux

L'exemple des consultations généralistes illustre notre approche de regroupement des actes de soins. Dans ce cas, nous avons intégré l'acte principal de consultation avec des majorations et des frais de déplacement, tout en rectifiant les incohérences dues aux variations orthographiques, telles que les différences entre majuscules et minuscules, garantissant ainsi une cohérence dans les données malgré les inévitables changements réglementaires et ajustements de gestion interne. Cette méthodologie de regroupement ne se limite pas à des actes homogènes, il englobe également des cas aux natures très diverses.

Par exemple, comme le montre le tableau 2.2 suivant, certains actes, bien que rarement effectués et très différents en nature, présentent des montants de remboursement presque identiques. Dans un souci de simplification et d'efficacité de notre base de modélisation, nous optons pour un regroupement de ces actes. Cette approche méthodique et pragmatique, tout en conservant une précision analytique, permet de réduire la complexité de notre modèle et d'assurer une gestion plus aisée des données.

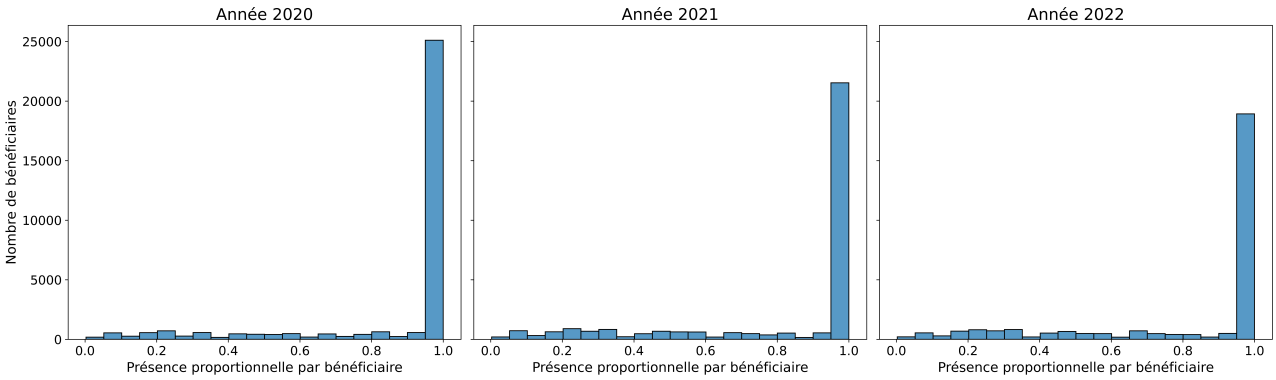
TABLE 2.2 – Exemple de regroupement des actes de soins - Spécialités Hors Nomenclature

Regroupement des actes	**Libellé d'acte de soins**
Spécialités Hors Nomenclature	Médecine alternative Acupuncture Consultation Psychologue

Calibration de nombre des soins de santé selon la présence au prorata

Typiquement, les contrats collectifs de santé complémentaire débutent le 1er janvier, fournissant ainsi une vision complète sur l'année. Cependant, des événements tels que des recrutements, des démissions, ou des changements de statut familial, tels que les mariages ou les naissances, peuvent altérer la période de couverture effective d'un bénéficiaire. Le graphique 2.4 ci-dessous présente la répartition proportionnelle de la durée de couverture des bénéficiaires pour la période de 2020 à 2022. Il ressort que, pour la majorité d'entre eux, les données couvrent des années complètes. Cependant, il existe des cas où les informations sont limitées à des périodes partielles de l'année.

FIGURE 2.4 – Répartition proportionnelle des présences des bénéficiaires de 2020 à 2022



En croisant les données des tables des prestations et des bénéficiaires, nous extrayons le nombre de soins de santé effectués et les coûts associés pour la période d'affiliation effective. Toutefois, pour établir une fréquence annuelle pertinente pour chaque regroupement de soins de santé, une calibration de nombre des soins de santé selon la présence au prorata des bénéficiaire est nécessaire. Cette démarche est essentielle pour une analyse précise, étant donné les variations dans la durée de couverture des bénéficiaires au sein des contrats collectifs de santé complémentaire.

Prenons l'exemple d'une consultation généraliste observée sur une période de 3 mois. Ici, le nombre "réel" de soins sur une année pourrait varier de 2 à 4. Dans ce cas, nous procédons à une calibration avant de calculer la fréquence annuelle. Cependant, pour d'autres types de soins, comme une hospitalisation chirurgicale ou l'achat de lunettes, une telle calibration n'est pas effectuée, basée sur notre expérience et les règles de remboursement.

Pour atténuer l'effet de la présence incomplète de certains assurés sur la modélisation de la fréquence, nous classons les soins de santé en trois catégories selon leur nature et les règles de remboursement en cas de durée de couverture incomplète : pas de calibration, calibration semestrielle et calibration trimestrielle. Le tableau 2.3 suivant illustre des exemples pour chacune de ces catégories. Cette segmentation nous permet de traiter les données de manière plus cohérente et de limiter les biais

dus à la présence incomplète, en tenant compte des variations dans la durée de couverture et la nature des soins.

TABLE 2.3 – Exemple de calibration de nombre des soins de santé

Méthode de calibration	**Regroupement d'actes de soins**
Pas de calibration	Forfait maternité Prothèses auditives Vaccin non remboursé ...
Calibration semestrielle	Échographie Orthodontique Pharmacie 30% ...
Calibration trimestrielle	Pharmacie 65% Kinésithérapie Consultation spécialiste ...

2.3 Analyse descriptive des données

Dans ce chapitre, dédié à la préparation des données pour la modélisation, une étape cruciale est l'analyse descriptive des données. Cette étape préliminaire est fondamentale, car elle permet de se familiariser avec les caractéristiques intrinsèques des données avant de procéder à la modélisation proprement dite.

L'analyse descriptive offre un aperçu complet des distributions statistiques des données, y compris leurs valeurs minimales et maximales. Cette compréhension détaillée est indispensable, en particulier dans le contexte de notre étude, où l'objectif est la tarification a posteriori utilisant des modèles bayésiens. Une telle approche nécessite une bonne connaissance sur la distribution des fréquences et des coûts des soins de santé.

En examinant ces distributions statistiques, nous pouvons non seulement évaluer la forme générale et les tendances des données, mais aussi identifier le type de distribution le plus approprié pour la loi a priori dans nos modèles bayésiens. Cette sélection judicieuse de la loi a priori est importante pour la précision et la validité de nos modèles prédictifs.

En outre, les graphiques de distribution de fréquence et de coût par regroupement d'actes de soins sont des outils visuels efficaces. Ils ne se limitent pas à illustrer la distribution générale des données, mais jouent également un rôle essentiel dans l'identification des valeurs aberrantes. Ces anomalies, si elles ne sont pas correctement traitées, peuvent induire des biais significatifs dans le modèle et fausser les résultats de la modélisation. Par conséquent, une attention particulière doit être accordée à ces points de données lors de l'analyse préparatoire, afin d'assurer la robustesse et la fiabilité de nos modèles de tarification.

Distribution de fréquence des soins de santé

La fréquence des actes de santé représente un élément déterminant dans l'élaboration de la tarification en santé. Chaque type et nature d'acte de soins de santé possède une distribution qui lui est propre, reflétant des modèles de consommation et des besoins en soins variés. Face à la diversité et à l'abondance des actes de soins de santé, même après un processus rigoureux de regroupement, il reste impossible de les représenter tous de manière exhaustive.

Tenant compte de ces considérations, notre analyse se concentrera sur la présentation et l'interprétation de cinq distributions représentatives de regroupements d'actes de soins. Chacun de ces regroupements a été sélectionné pour illustrer des natures et des comportements distincts, offrant ainsi une perspective nuancée et informative sur la distribution de fréquence des soins de santé. Cette démarche vise non seulement à simplifier notre approche analytique, mais aussi à mettre en lumière les tendances et les particularités significatives qui influencent directement la tarification des soins de santé.

Fréquence - Analyses médicales

Le graphique 2.5 illustre les fréquences annuelles des actes d'analyses médicales effectuées en médecine de ville sur une période de quatre ans, de 2019 à 2022. Pour optimiser la lisibilité, notamment pour les fréquences d'actes élevées, les fréquences des bénéficiaires qui n'ont effectué aucune analyse médicale ne sont pas représentées dans les barres du graphique. Ces données sont cependant calculées et présentées de manière synthétique dans le tableau situé en haut à droite du graphique.

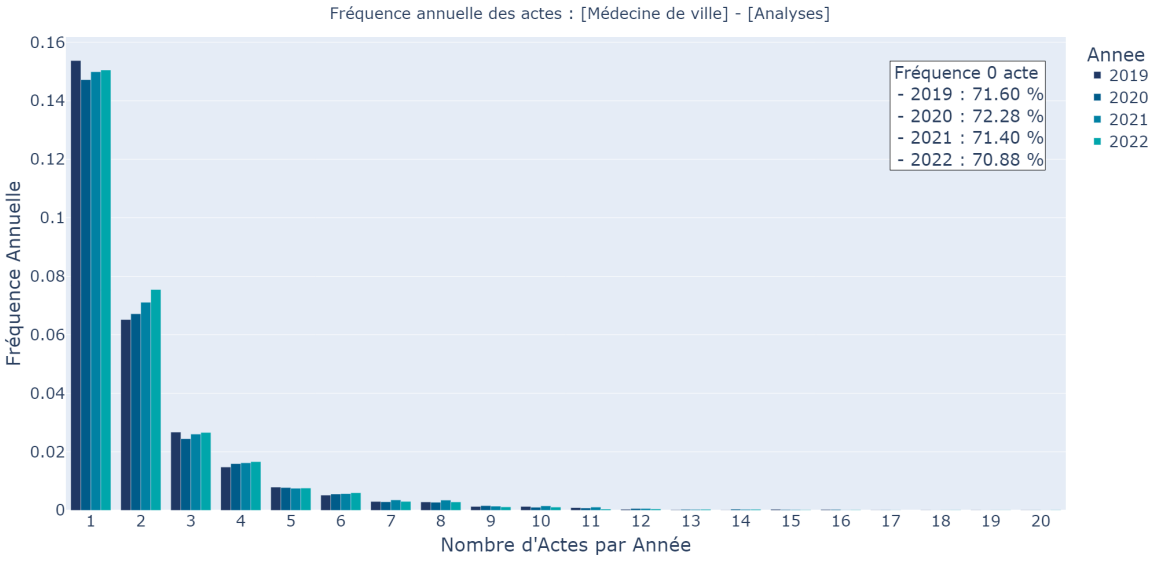


FIGURE 2.5 – Répartition des fréquences des analyses médicales

Une première observation concerne les fréquences d'absence d'actes, où l'on remarque que l'année 2020 se distingue par la fréquence la plus haute. Cette particularité s'explique par la pandémie de Covid-19, durant laquelle les tests de dépistage ont été pris en charge intégralement par la Sécurité sociale. De plus, les mesures de confinement ont entraîné une diminution des autres analyses médicales comparativement aux autres années.

Concernant les fréquences des actes d'analyses médicales, nous constatons que près de 15% des individus en réalisent une fois par an, un pourcentage qui se maintient relativement constant sur les

quatre années observées. Il est également notable que 14% de la population étudiée effectue plus d'une analyse médicale par an.

La distribution observée suggère une similitude avec les lois discrètes telles que la loi de Poisson ou la loi binomiale négative, qui sont souvent utilisées pour modéliser le nombre d'événements sur une période donnée. Cependant, la présence non négligeable de fréquences élevées d'actes (supérieures à 5) nous incite à la prudence. L'application directe d'une loi de Poisson ou binomiale négative pourrait en effet conduire à une surévaluation des fréquences de 0 et 1 acte et à une sous-estimation pour les fréquences plus élevées. Cette considération doit guider le choix du modèle statistique approprié pour une représentation fidèle de la réalité et pour éviter les biais dans l'estimation des fréquences d'actes médicaux.

Fréquence - Consultation généraliste

Le graphique 2.6 présenté ci-dessous détaille les fréquences annuelles des consultations généralistes entre 2019 et 2022. Suivant la même logique que le graphique précédent concernant les analyses médicales, la fréquence de 0 acte n'est pas affichée directement dans les barres du graphique pour permettre une meilleure distinction des fréquences des nombres plus élevés d'actes. Néanmoins, ces données sont calculées et clairement résumées dans le tableau situé en haut à droite du graphique.

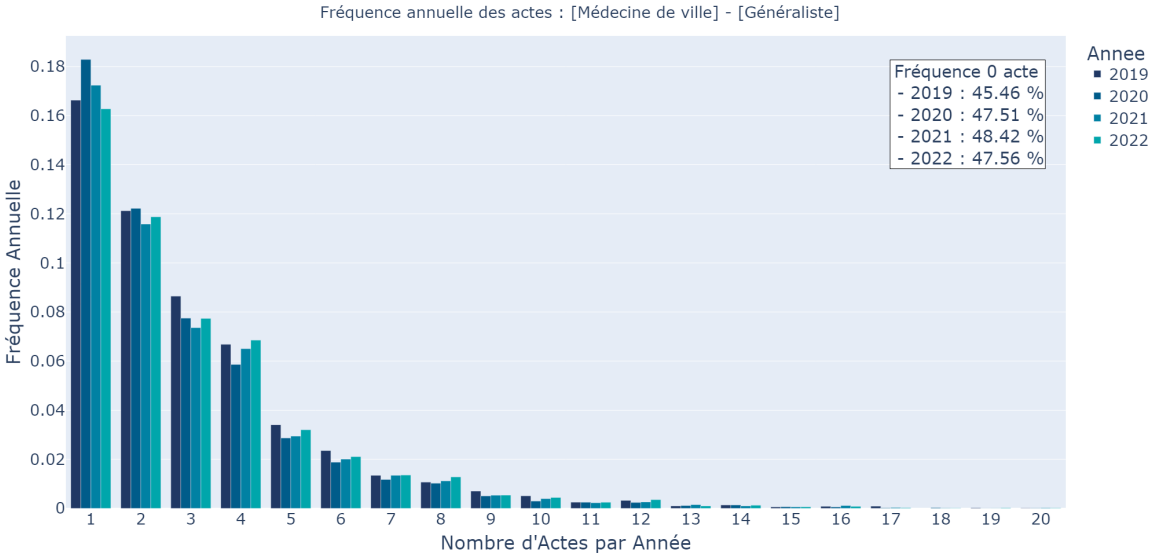


FIGURE 2.6 – Répartition des fréquences des consultations généralistes

L'observation des données révèle une décroissance de la fréquence avec l'augmentation du nombre de consultations. Parmi les individus ayant réalisé au moins une consultation annuelle, qui représentent plus de la moitié de la population, la moyenne annuelle était de 3,1 consultations par personne. Cette moyenne, spécifique aux personnes consultées, varie légèrement en raison de l'impact de la pandémie de Covid-19, avec des moyennes de 3,0 pour 2021 et de 2,9 pour 2020. Un autre point d'intérêt est que près de 10% des bénéficiaires ont réalisé plus de trois consultations généralistes par an.

La forme de cette distribution suggère que les fréquences des consultations généralistes ne suivent pas un modèle qui serait adéquatement représenté par des lois statistiques simples telles que la loi de Poisson ou la loi binomiale négative. Cela indique le besoin d'adopter des modèles plus complexes pour

capturer la distribution réelle des données. Dans le prochain chapitre, nous explorerons les modèles de mélange comme solution potentielle pour modéliser avec plus de précision ces distributions.

Fréquence - Kinésithérapie

Le graphique 2.7 affiché ci-dessous illustre les fréquences annuelles des séances de kinésithérapie pour les années allant de 2019 à 2022. De manière similaire aux graphiques précédents sur les analyses médicales et les consultations généralistes, la fréquence de 0 acte n'est pas représentée dans les barres, mais ces données sont précisément calculées et affichées dans le résumé en haut à droite du graphique. Afin de simplifier l'analyse, nous avons délibérément omis les données représentant plus de 90 actes, bien que dans la réalité, le nombre d'actes puisse atteindre plus de 160.

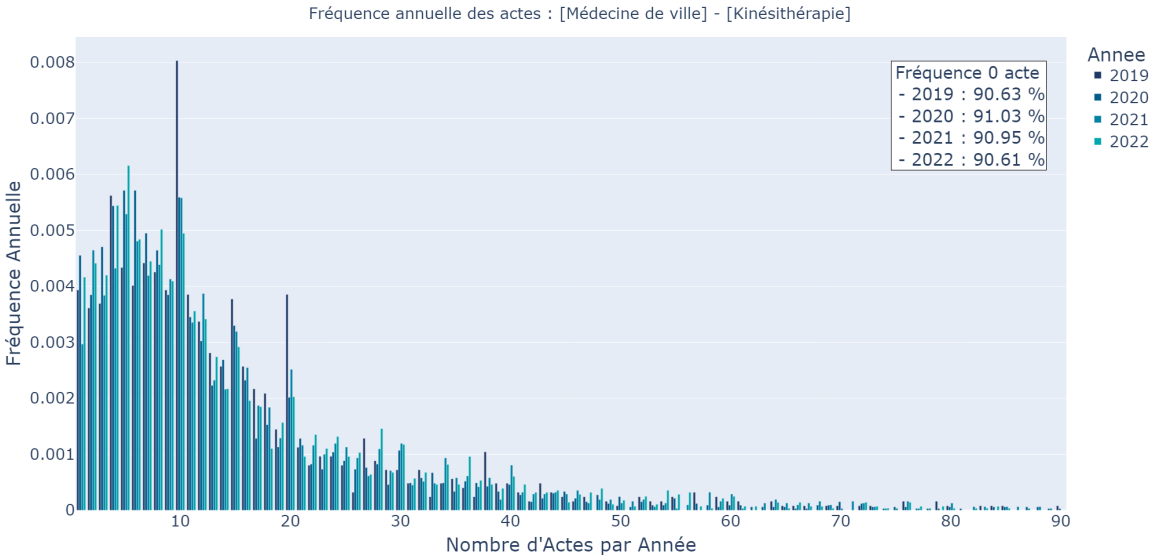


FIGURE 2.7 – Répartition des fréquences de la kinésithérapie

L'analyse de la kinésithérapie révèle un modèle distinct : seulement environ 10% de la population concernée par ces soins, mais avec une moyenne annuelle de séances relativement élevée, s'élevant à 15,5 (avec une diminution à 14,5 en 2020 et une augmentation à 16,3 en 2021, variations probablement attribuables aux effets de la pandémie de Covid-19). Contrairement aux analyses médicales et aux consultations généralistes, les fréquences pour un nombre d'actes de 1 à 3 ne sont pas les plus courantes, ce qui est cohérent avec la pratique de la kinésithérapie où plusieurs séances sont souvent nécessaires pour obtenir un bénéfice thérapeutique.

En tenant compte de ces spécificités, il est clair qu'une loi statistique simple ne suffirait pas à modéliser correctement cette distribution. Les modèles à inflation de zéros, qui sont une forme spécifique de modèles de mélange, pourraient en effet fournir une meilleure approximation pour ces données. Ce type de modèle prend en compte la probabilité élevée de non-usage (les zéros) tout en modélisant la distribution des nombres plus élevés de séances, ce qui semble être un outil adapté pour cette distribution particulière de fréquences en kinésithérapie.

Fréquence - Orthophonie et orthoptie

Le graphique 2.8 ci-dessous dépeint les fréquences annuelles des actes d'orthophonie et d'orthoptie sur la période de 2019 à 2022. À l'instar des autres graphiques, la fréquence de 0 acte est omise des barres pour une meilleure lisibilité et ces données sont mises en évidence dans le tableau en haut à droite.

Nous avons ici choisi de regrouper les actes d'orthophonie et d'orthoptie afin de réduire le nombre final de regroupements d'actes de soins. Contrairement au cas des consultations généralistes, il n'existe pas de lien direct entre l'orthophonie et l'orthoptie. Cependant, l'analyse des données historiques a montré que ces deux actes sont indépendants, peu fréquents, et que les montants de remboursement du régime complémentaire sont assez proches. Dans le but de simplifier le modèle final, nous avons donc décidé de les regrouper. Toutefois, pour une analyse plus précise, il serait nécessaire d'étudier ces deux actes séparément.

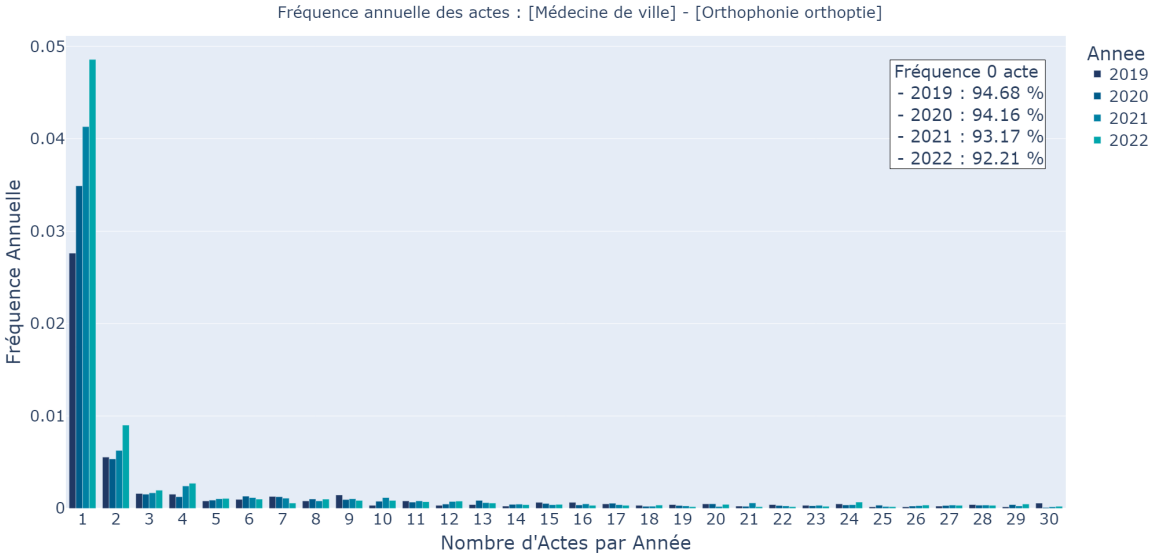


FIGURE 2.8 – Répartition des fréquences de l'orthoptie et de l'orthophonie

Dans le domaine de l'orthophonie et de l'orthoptie, la distribution fréquentielle démontre une configuration particulière. Ces services concernent une fraction réduite de la population mais il est évident qu'il existe une tendance à l'augmentation des fréquences lorsque le nombre d'actes est supérieur à 0. De manière frappante, le nombre d'actes unitaires est prédominant, suggérant l'existence de consultations uniques sans suivi thérapeutique. En revanche, dès le seuil de 5 actes franchi, la distribution semble indiquer des parcours de soins ou des suivis récurrents.

Semblablement aux observations précédentes, une loi de comptage simple ne serait pas adaptée pour modéliser cette distribution. Il semble nécessaire d'adopter une approche plus complexe, en combinant différentes distributions pour mieux refléter la réalité des données. Cette méthodologie, qui sera exposée dans les sections suivantes, vise à fournir une estimation plus fidèle de la distribution en tenant compte de la diversité des schémas de soins en orthophonie et orthoptie.

Fréquence - Chirurgie de l'œil au laser

Le graphique 2.9 ci-dessous montre les fréquences annuelles des actes de chirurgie de l'œil au laser pour les années 2019 à 2022. Comme pour les autres données de santé précédemment analysées, la fréquence des cas sans acte chirurgical n'est pas représentée directement sur les barres du graphique, mais ces informations sont néanmoins compilées et affichées dans le coin supérieur droit pour fournir une vue complète.

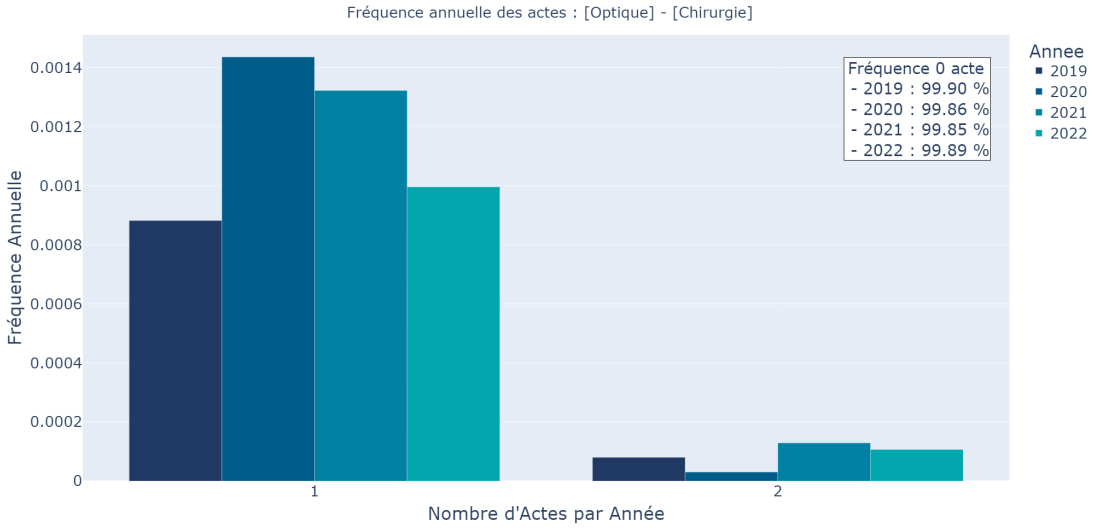


FIGURE 2.9 – Répartition des fréquences de la chirurgie optique

La chirurgie au laser pour les yeux est une procédure médicale particulière, comparable aux équipements pour personnes handicapées ou aux prothèses auditives, en ce sens que les occurrences sont relativement rares, mais associées à des coûts élevés. La plupart des actes recensés ici, indiqués par le chiffre 1, peuvent représenter une procédure bilatérale, c'est-à-dire effectuée sur les deux yeux lors d'une même session. Le fait de voir également un nombre d'actes représentés par le chiffre 2 n'est pas considéré comme une anomalie, cela indique plutôt que certains patients ont subi deux interventions distinctes, probablement une pour chaque œil.

Les méthodes classiques de modélisation basées sur des données historiques, telles que les modèles linéaires généralisés *(GLM)*, peuvent rencontrer des difficultés si les actes n'ont pas été effectués dans le passé, car l'absence d'antécédents annulerait potentiellement l'effet prédit. Cependant, les modèles bayésiens, pourvu d'une loi a priori bien choisie, offrent une solution à ce problème. Ils permettent d'intégrer l'information sur l'absence d'actes passés d'une manière qui ajuste et affûte les prédictions pour l'avenir, fournissant une vue plus complète et plus nuancée des probabilités d'occurrence d'actes chirurgicaux.

Distribution des coûts des soins de santé

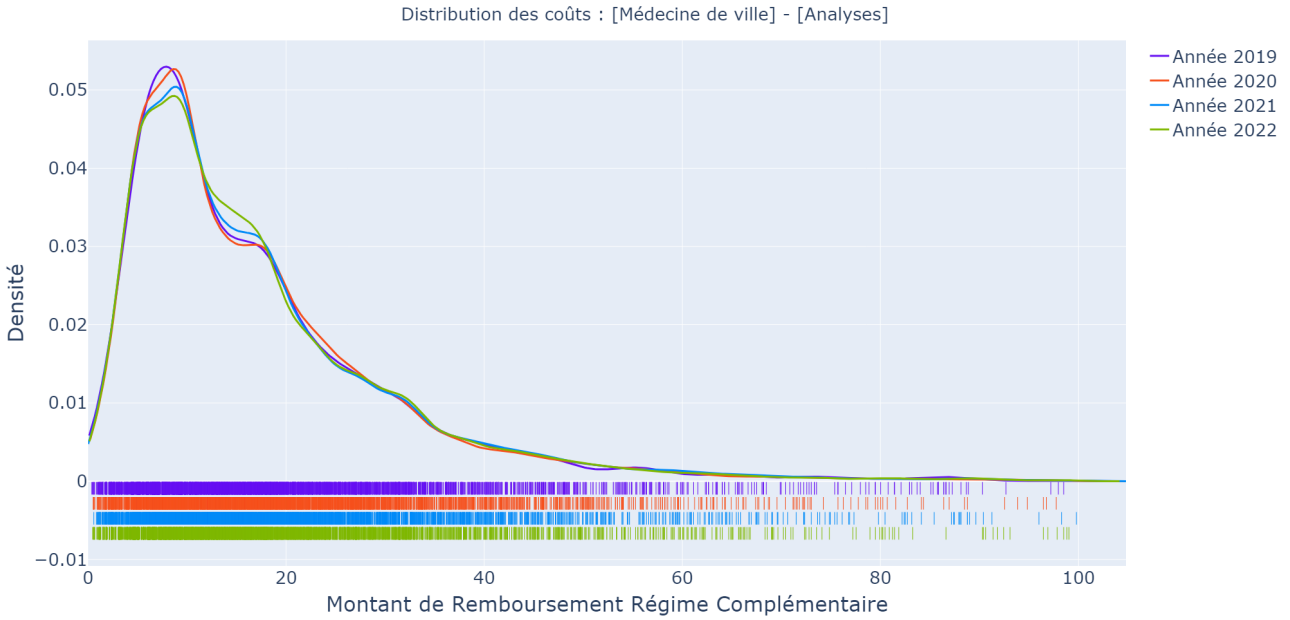
L'analyse des coûts associés aux actes de santé constitue un pilier fondamental dans l'élaboration de modèles tarifaires précis et efficaces dans le domaine de la santé. La nature et le type de chaque acte de soins de santé engendrent des distributions de coûts hétérogènes, reflétant la complexité et la diversité des interventions médicales. En raison de la vaste gamme d'actes de soins disponibles, il s'avère impraticable de les examiner tous individuellement, même après un regroupement minutieux.

Dans ce contexte, notre étude s'attachera à examiner et à interpréter cinq distributions clés, chacune représentant un regroupement d'actes de soins aux caractéristiques et comportements distincts. Cette sélection vise à fournir un aperçu significatif des différentes gammes de coûts rencontrées, tout en soulignant les particularités qui caractérisent chaque catégorie. Cette approche ciblée permet non seulement d'appréhender la complexité inhérente à la tarification des soins de santé, mais aussi d'identifier des tendances clés et des comportements de coût significatifs. Notre objectif est de fournir une compréhension plus nuancée et détaillée des structures de coûts, un aspect crucial pour une tarification équilibrée et efficace dans le secteur de la santé.

Analyse des remboursements - Analyses médicales

Le graphique 2.10 ci-dessous illustre les distributions annuelles des remboursements effectués par le régime complémentaire pour des analyses médicales entre 2019 et 2022. Pour simplifier l'analyse, les données comprenant des montants de remboursement excédant 100 euros ont été exclues.

FIGURE 2.10 – Répartition des remboursements des analyses médicales



Les courbes de densité, déclinées en plusieurs couleurs, tracent les distributions des montants de remboursement année par année. L'axe horizontal représente les montants remboursés par le régime complémentaire, tandis que l'axe vertical mesure la densité, pouvant être perçue comme la probabilité estimée d'occurrence d'un montant donné. Les pics sur ces courbes signalent les tranches de remboursement les plus récurrentes.

Au bas du graphique, les barres colorées correspondent aux données individuelles pour chaque année, assorties aux couleurs des courbes de densité. Cette représentation graphique fournit un aperçu de la dispersion des montants remboursés, soulignant où les données se concentrent.

Notamment, la distribution affiche une multimodalité, avec un pic prédominant situé entre 5 et 10 euros, suggérant que cette gamme de remboursement est la plus fréquemment rencontrée. D'autres pics, par exemple entre 15 et 20 euros, peuvent indiquer l'existence de différents services d'analyses

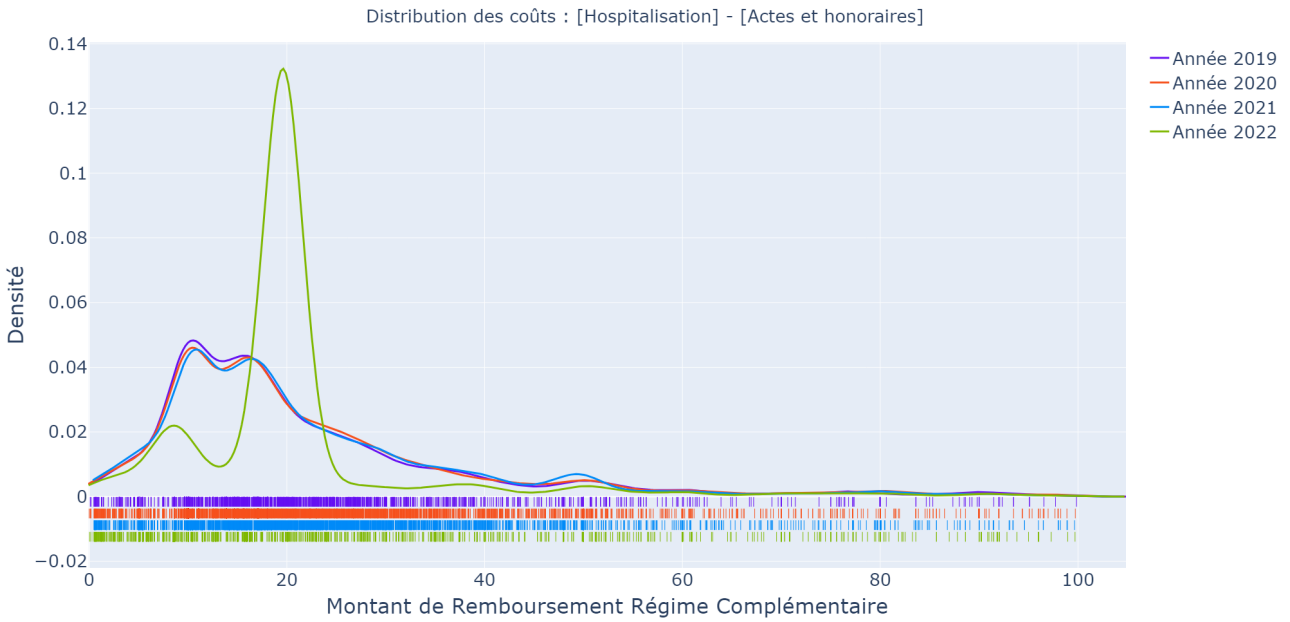
médicales ayant des coûts variés.

En comparant les distributions sur les quatre années, on observe une stabilité relative des montants de remboursement, indiquant peu de variation annuelle pour les analyses médicales, contrairement à d'autres types de soins qui pourraient présenter des fluctuations plus marquées d'une année à l'autre. Ce constat met en évidence l'importance de considérer la spécificité des services lors de l'analyse des distributions et des tendances de remboursement en santé.

Analyse des remboursements - Services hospitaliers

Le graphique 2.11 ci-dessous illustre la distribution des remboursements du régime complémentaire pour les actes et honoraires hospitaliers de 2019 à 2022. Pour rendre les données plus digestes, les montants de remboursement supérieurs à 100 euros n'ont pas été inclus, bien que dans la réalité, ces montants puissent excéder 4000 euros annuellement.

FIGURE 2.11 – Répartition des remboursements des services hospitaliers



Contrairement à la distribution pour les analyses médicales, nous remarquons un changement notable en 2022 avec un pic autour de 20 euros. Cette variation pourrait être influencée par l'inclusion du Forfait Patient Urgences *(FPU)* dans l'ensemble des données. Le FPU est une charge fixe de 19,61 euros mise en place à partir du 1er janvier 2022 pour les patients qui se rendent aux urgences et qui ne sont pas hospitalisés par la suite. Cette mesure vise à informer les patients du coût de leur passage aux urgences immédiatement après leur visite, plutôt que d'attendre plusieurs semaines ou mois pour recevoir une facture. Le FPU est remboursé intégralement par la mutuelle ou la complémentaire santé.

En observant la distribution, nous identifions divers sommets distincts, ce qui suggère l'existence de divers types de services hospitaliers engendrant différents coûts. Ce phénomène reflète la complexité et la diversité des actes et honoraires hospitaliers, qui englobent une large gamme de procédures et de services, y compris, mais sans s'y limiter, le FPU.

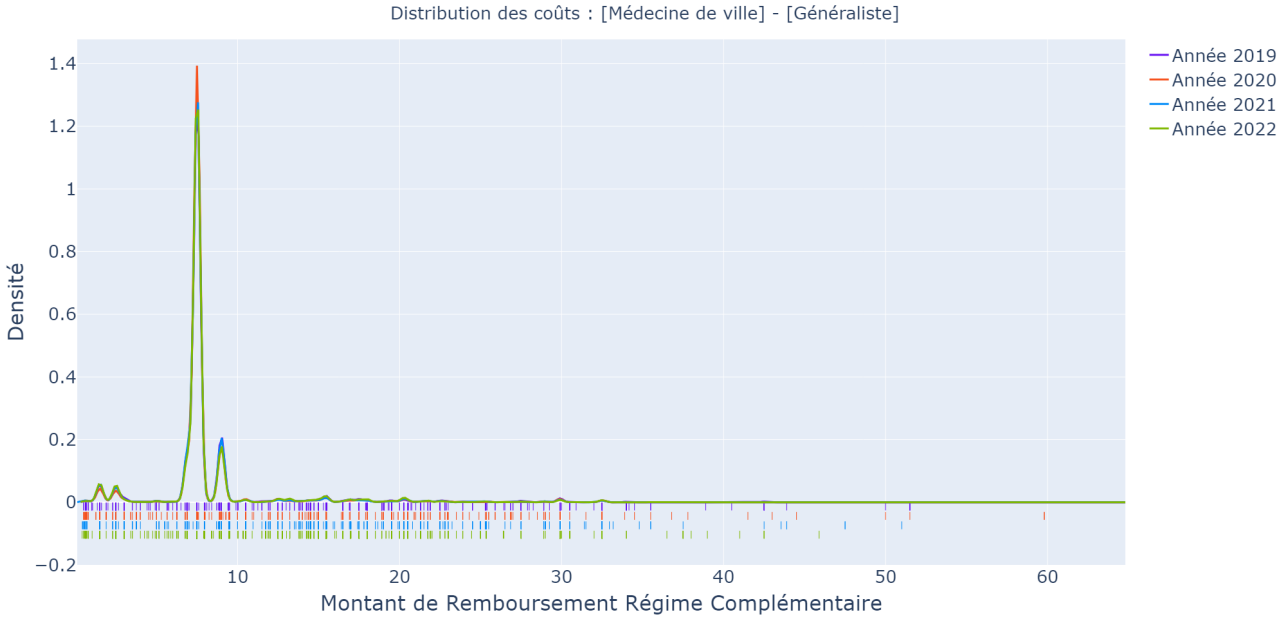
Pour chaque année représentée, la queue de la distribution s'étale à droite des pics, indiquant

la décroissance des fréquences des remboursements à mesure que les montants augmentent, ce qui suggère que les montants très élevés de remboursement sont rares dans l'ensemble des données. En d'autres termes, il y a une plus grande probabilité d'observer des remboursements plus faibles et une probabilité décroissante pour les montants plus élevés. Cela indique également qu'il y a une variété de remboursements qui sont moins courants mais toujours possibles. Cette longue queue peut être caractéristique d'une distribution avec une variance relativement élevée.

Analyse des remboursements - Consultations généralistes

Le graphique 2.12 montre la distribution des coûts remboursés par le régime complémentaire pour des consultations généralistes de 2019 à 2022. La distribution est clairement multimodale, avec plusieurs pics marquants différentes situations de remboursement.

FIGURE 2.12 – Répartition des remboursements des consultations généralistes



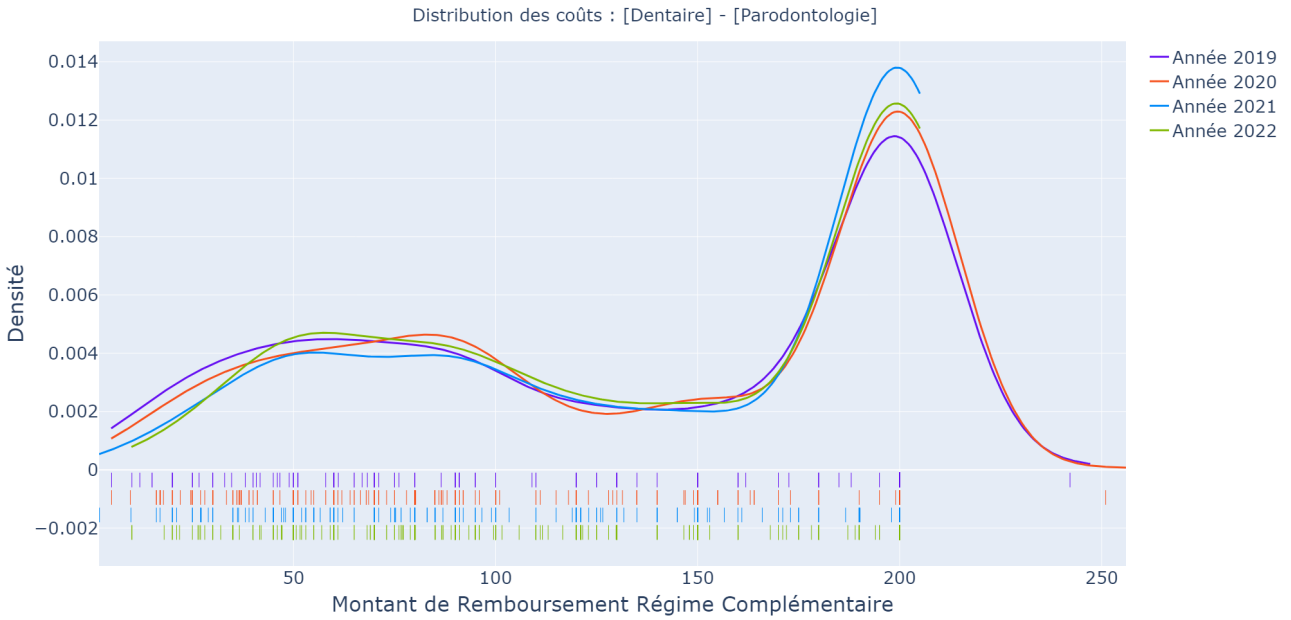
Le pic le plus saillant, situé à de 7,5 euros, est le remboursement standard pour une consultation générale, après prise en compte du remboursement de la Sécurité sociale et du forfait patient. Le pic à 9 euros représente le remboursement standard de 7,5 euros augmenté d'une majoration de 1,5 euro pour dépassement d'honoraires, indiquant ainsi une consultation standard avec un supplément précis pour des honoraires légèrement supérieurs. Les pics moins élevés autour de 2 euros pourraient correspondre aux consultations dans le cadre du régime local Alsace-Moselle, qui offre des conditions de remboursement spécifiques.

La queue de la distribution, qui s'étend jusqu'à 60 euros, illustre la variabilité des coûts pour des actes complémentaires ou des situations exceptionnelles comme les déplacements, les consultations les jours fériés ou de nuit.

Analyse des remboursements - Soins dentaires non remboursés

Le graphique 2.13 ci-dessous représente la distribution des remboursements effectués par le régime complémentaire pour des soins dentaires non remboursés par la Sécurité sociale, sur la période de 2019 à 2022.

FIGURE 2.13 – Répartition des remboursements des soins dentaires non remboursés



L'analyse graphique révèle une distribution tronquée à 200 euros, ce qui correspond au niveau de garantie maximal attribué par an et par bénéficiaire pour ce type de soins. Bien que ce plafond de garantie soit généralement respecté, on remarque la présence d'individus en 2019 et 2020 avec des remboursements qui excèdent ce plafond, ce qui pourrait indiquer des anomalies ou des exceptions spécifiques. Pour les besoins de modélisation, et compte tenu du nombre très limité de ces dépassements, il serait raisonnable d'exclure ces points de données afin de maintenir la cohérence de l'analyse et d'éviter toute distorsion dans les calculs.

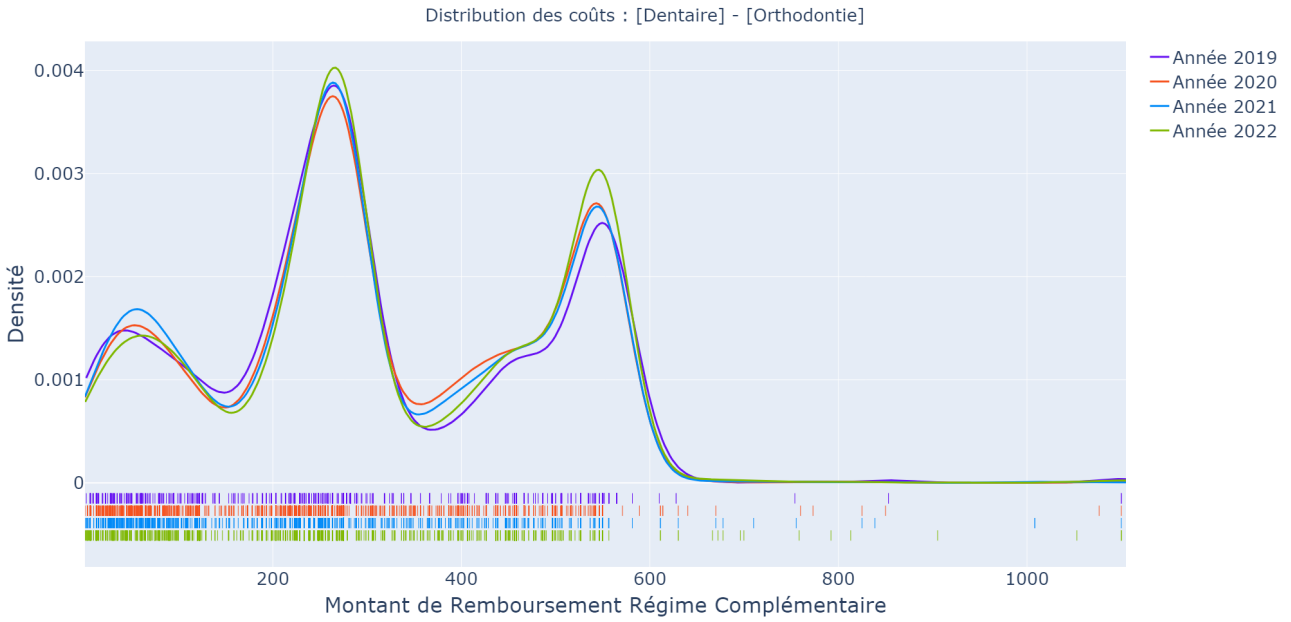
Les variations subtiles observées dans les courbes annuelles suggèrent que les montants de remboursement restent relativement stables d'année en année. Il y a des variations significatives dans la densité des montants de remboursement, avec des zones de densité plus élevée suivies de zones de densité plus faible.

Analyse des remboursements - Traitements d'orthodontie

Le graphique 2.14 dépeint les distributions des montants remboursés pour les traitements orthodontiques par le régime complémentaire de 2019 à 2022.

Nous observons trois pics majeurs dans les distributions. Le premier pic, qui est le plus bas, se situe autour de 60 euros, le second est aux alentours de 260 euros, et le troisième pic se trouve près de 550 euros. Ces pics pourraient représenter des traitements ou des séries de soins orthodontiques spécifiques qui sont fréquemment remboursés à ces niveaux de coûts. La queue de la distribution est allongée et s'étend vers les montants plus élevés, indiquant que bien que moins fréquents, il existe des

FIGURE 2.14 – Répartition des remboursements des traitements d'orthodontie



remboursements importants pour certains traitements orthodontiques coûteux.

Les distributions annuelles montrent des petites variations, ce qui pourrait suggérer de légères différences dans les types de traitements orthodontiques réalisés. Cependant, la forme générale des distributions reste similaire au fil des ans, ce qui indique une certaine stabilité dans les pratiques de remboursement. Les sommets des courbes pour chaque année sont relativement proches en hauteur, suggérant qu'il n'y a pas eu de changements drastiques dans les montants moyens remboursés ni dans la fréquence des différents montants de remboursement au cours des quatre années.

Conclusion

Dans ce deuxième chapitre intitulé "Préparation des données en vue de la modélisation", nous avons méticuleusement analysé et préparé les données, en établissant des hypothèses et en gérant la complexité inhérente à ces informations. Malgré leur nature complexe, les données recèlent des particularités et des patterns sous-jacentes exploitables. Notre processus rigoureux a inclus la collecte, le nettoyage, l'agrégation et le regroupement minutieux des données, notamment des actes de soins, en tenant compte de critères tels que la fréquence et les coûts. Nous avons également accordé une attention particulière à l'analyse descriptive, essentielle pour saisir les caractéristiques intrinsèques des données, y compris leurs bornes, qui sont cruciales dans le cadre de notre objectif de tarification a posteriori.

En s'appuyant sur ces données historiques préparées, le prochain chapitre abordera la modélisation à l'aide de modèles bayésiens. Cette approche nous permettra d'exploiter les nuances et les modèles cachés dans les données, en vue d'une tarification a posteriori précise. Cette étape représente un pont entre la préparation approfondie des données et leur application pratique dans le domaine complexe de l'actuariat, illustrant l'efficacité de l'analyse actuarielle lorsqu'elle est soutenue par une préparation adaptée et des méthodes de modélisation avancées.

Chapitre 3

Approche bayésienne en tarification a posteriori

3.1 Fondements de la tarification en assurance santé collective

La tarification en assurance santé collective repose sur l'utilisation de données démographiques généralement synthétiques. Plutôt que de fournir des informations détaillées pour chaque salarié, les entreprises communiquent habituellement des statistiques globales, souvent désignées comme démographie agrégée. Ces données synthétiques sont par la suite converties en informations plus détaillées grâce à des outils de tarification tels que Addactis Prévoyance Office. Ces outils emploient diverses méthodes de calcul pour répartir la population couverte en fonction de critères spécifiques tels que le sexe, la situation familiale et le nombre d'enfants.

Afin de calculer la prime pure moyenne pour chaque individu au sein d'une entreprise, on se base sur l'agrégation des prestations probables pour chaque ligne de garantie. Cette somme repose sur trois éléments fondamentaux, énumérés ci-dessous :

- **Coût moyen de la prestation par acte** : Calculé en prenant en compte les garanties offertes, les frais engagés et le remboursement potentiel de la Sécurité sociale.
- **Fréquence de consommation par acte** : Évaluée en fonction de la fréquence à laquelle les différents actes médicaux sont utilisés, ajustée selon les caractéristiques démographiques de la population couverte.
- **Correctifs** : Utilisés pour ajuster la consommation d'un acte ou d'un individu. Il existe plusieurs types de correctifs, incluant l'âge, le sexe, la catégorie socio-professionnelle *(CSP)*, la localisation géographique, la classe de risque, le montant et d'autres éléments pertinents.

Tarification a priori

La tarification a priori en assurance santé collective se base sur l'évaluation des risques avant la réalisation des sinistres. Dans cette approche, l'assureur utilise des données démographiques et statistiques historiques pour déterminer la prime d'assurance. Cette méthode est généralement employée pour les nouveaux contrats ou pour des groupes pour lesquels on ne dispose pas d'un historique de sinistralité. Elle permet ainsi d'établir une prime de base qui reflète le risque moyen prévu pour le groupe ciblé.

Le recours à la tarification a priori se justifie par la nécessité de fixer des primes avant que l'expérience réelle ne soit connue. Elle repose sur des hypothèses actuarielles concernant la fréquence et le coût des sinistres, ajustées en fonction de variables telles que l'âge, la zone géographique, la composition de la famille, et d'autres facteurs de risque connus. Les assureurs établissent des grilles tarifaires en utilisant des modèles tels que les modèles multiplicatifs ou les modèles linéaires généralisés

(GLM), qui prennent en compte ces variables pour déterminer les primes pour les nouveaux assurés dont l'historique de sinistres est inconnu. L'avantage de cette méthode est qu'elle permet de définir une structure tarifaire stable et prévisible, qui est facilement compréhensible pour les assurés.

Toutefois, cette approche présente l'inconvénient de ne pas prendre en compte la sinistralité individuelle réelle. Elle peut donc mener à une situation où des assurés moins risqués subventionnent indirectement ceux qui présentent un risque plus élevé. De plus, elle ne récompense pas les comportements vertueux en matière de gestion des risques santé, puisque tous les membres du groupe sont initialement tarifés de la même manière.

Tarification a posteriori

La tarification a posteriori, en revanche, est une approche dynamique qui ajuste les primes en fonction de l'expérience de sinistralité réelle des assurés. Elle s'appuie sur les données de sinistralité individuelles ou de groupe collectées au fil du temps. Cette méthode est particulièrement pertinente pour les contrats en renouvellement ou pour les groupes ayant un historique de sinistralité suffisamment riche pour permettre une analyse détaillée.

Dans cette approche, les primes peuvent être ajustées en fonction de la performance relative d'un groupe ou d'un individu par rapport aux attentes initiales. Cela encourage une meilleure gestion des risques de santé, car les assurés ont un intérêt financier direct à maintenir ou à améliorer leur état de santé ou leurs comportements en matière de prévention et de soins.

La tarification a posteriori offre l'avantage d'être plus équitable, puisqu'elle reflète mieux le principe de la mutualisation des risques homogènes. Contrairement à une discrimination tarifaire basée sur l'état de santé, cette approche permet plutôt d'évaluer le niveau de risque des entreprises et de mutualiser les risques entre celles ayant un profil de risque similaire. Cela évite que les entreprises à faible risque ne subventionnent excessivement celles à risque plus élevé, ce qui pourrait entraîner des résiliations de bons risques et, par conséquent, la dégradation du portefeuille. Cependant, la tarification a posteriori peut être perçue comme complexe par les assurés, car elle peut entraîner des variations significatives des primes d'une année sur l'autre. De plus, elle peut introduire une certaine volatilité dans la structure tarifaire et nécessite des systèmes d'information et de gestion des données robustes.

En somme, la tarification a priori offre une approche plus stable et prévisible, bien adaptée pour les nouveaux assurés ou pour les groupes sans historique de sinistralité, tandis que la tarification a posteriori permet une personnalisation et une équité tarifaire accrues, idéales pour les groupes avec un historique de sinistralité. Pour les deux types de tarification, des modèles mathématiques et statistiques complexes sont utilisés pour estimer le risque et fixer les primes de manière appropriée. Ces modèles peuvent inclure des approches bayésiennes, qui permettent d'ajuster les estimations en intégrant de nouvelles informations au fur et à mesure qu'elles deviennent disponibles.

3.2 Modèle bayésien : une synthèse entre données et expertise

L'approche bayésienne dans la tarification a posteriori en assurance apporte une perspective différente par rapport aux méthodes traditionnelles. Elle permet d'intégrer de manière fluide les informations passées des assurés avec des préjugés ou connaissances antérieures dans un modèle prédictif robuste et évolutif.

Le modèle bayésien

Dans le contexte actuariel, le modèle bayésien est particulièrement pertinent pour la tarification a posteriori où les estimations de risque doivent être mises à jour de manière dynamique. Les éléments suivants sont essentiels à la compréhension de cette approche :

- **Distribution a priori** : La distribution a priori, qui encapsule les croyances antérieures sur les paramètres du modèle avant l'examen des données, notée $p(\theta)$, représente nos croyances initiales sur le paramètre θ avant l'observation des données.

- **Vraisemblance** : La vraisemblance, notée $p(Y|\theta)$, évalue la probabilité des données observées Y étant donné un paramètre θ.

- **Distribution a posteriori** : La distribution a posteriori, $p(\theta|Y)$, est la mise à jour de nos croyances après avoir pris en compte les nouvelles données Y. Elle est calculée en utilisant la formule de Bayes :

$$p(\theta|Y) = \frac{p(Y|\theta) \cdot p(\theta)}{p(Y)}$$

  Dans cette formule, $p(Y)$ est la probabilité marginale ou l'évidence, qui agit comme un facteur normalisant s'assurant que la distribution a posteriori soit une distribution de probabilité valide.

- **Probabilité marginale** : La probabilité marginale, $p(Y)$, est l'intégrale de la vraisemblance sur tous les paramètres possibles, pondérée par la distribution a priori. Cette intégration est exprimée mathématiquement comme suit :

$$p(Y) = \int_\theta p(Y|\theta) \cdot p(\theta)\, d\theta$$

Alors que le modèle bayésien offre un cadre puissant pour la mise à jour des croyances en présence de nouvelles données, sa mise en œuvre pratique rencontre souvent des défis computationnels. Un de ces défis est le calcul de la probabilité marginale $p(Y)$, qui joue le rôle de facteur normalisant dans la distribution a posteriori. Il est courant en pratique de reformuler la formule de Bayes sous une forme proportionnelle simplifiée. Cette reformulation met en évidence la relation directe entre les distributions a posteriori, la vraisemblance et a priori, sans la nécessité immédiate de calculer le dénominateur complexe. Mathématiquement, cette formule proportionnelle est souvent exprimée comme suit :

$$p(\theta|Y) \propto p(Y|\theta) \cdot p(\theta)$$

Cette expression illustre que la distribution a posteriori est proportionnelle au produit de la vraisemblance des données observées et de la distribution a priori des paramètres. Cette approche nous libère du fardeau du calcul de l'intégrale complète, permettant une exploration plus pratique de l'espace des paramètres, ce qui est particulièrement utile lorsque nous nous tournons vers des techniques numériques avancées comme les Méthodes de Monte-Carlo par chaînes de Markov *(MCMC)* pour résoudre le problème.

Méthodes de Monte-Carlo par chaînes de Markov

La probabilité marginale $p(Y)$ est une intégrale sur tous les paramètres possibles θ, et peut être très difficile à calculer directement, surtout lorsque l'espace des paramètres est vaste ou complexe. Le défi réside dans le fait que pour la plupart des problèmes pratiques, une expression fermée pour $p(Y)$ n'est pas disponible, ce qui nous contraint à recourir à des méthodes numériques pour approximer cette intégrale.

Face à cette complexité, les méthodes Monte-Carlo par chaînes de Markov émergent comme un outil précieux. Ces techniques permettent d'échantillonner efficacement à partir de la distribution a posteriori sans nécessiter le calcul explicite de $p(Y)$. En générant des échantillons qui suivent la distribution a posteriori, les méthodes MCMC nous permettent de construire des estimations des paramètres et de faire des prédictions, tout en incorporant l'incertitude inhérente à nos estimations.

Les méthodes MCMC reposent sur la construction d'une chaîne de Markov, qui est une séquence de variables aléatoires où la probabilité de chaque événement dépend uniquement de l'état précédent. L'objectif est de créer une chaîne de Markov dont la distribution à long terme, vers laquelle la chaîne converge après de nombreuses itérations, correspond à la distribution a posteriori que l'on cherche à échantillonner.

Processus d'échantillonnage MCMC

- **Initialisation** : On commence par choisir un point de départ arbitraire dans l'espace des paramètres, souvent basé sur la distribution a priori ou une estimation informée.
- **Échantillonnage itératif** : À chaque étape, un nouvel échantillon est généré en fonction de l'échantillon précédent, suivant une certaine proposition de distribution. Cette proposition est souvent choisie pour faciliter le calcul et assurer une bonne exploration de l'espace des paramètres. :
- **Critère d'acceptation (méthode de Metropolis-Hastings)** : Chaque nouvel échantillon proposé est accepté avec une probabilité déterminée par le ratio de la vraisemblance des nouveaux et anciens échantillons, ajusté par le ratio des distributions a priori correspondantes. Si l'échantillon n'est pas accepté, la chaîne reste dans l'état précédent.
- **Convergence** : Avec suffisamment d'itérations, la chaîne de Markov atteint un état de convergence où les échantillons sont considérés comme étant tirés de la distribution a posteriori.

Distributions a priori et lois statistiques usuelles

L'approche bayésienne en statistique repose sur un choix judicieux des distributions a priori. Ces distributions représentent nos connaissances ou croyances préalables sur les paramètres du modèle avant de prendre en compte les données observées. Leur sélection influence significativement les résultats des analyses bayésiennes et, par conséquent, requiert une attention particulière.

En effet, nous disposons souvent d'informations pertinentes sur les paramètres d'intérêt, telles que des contraintes logiques, physiques, ou des données historiques. Ces informations guident le choix des distributions a priori informées. Bien que discrètes, elles apportent une valeur ajoutée significative à l'analyse. Par exemple, un paramètre spécifique peut être naturellement contraint à des valeurs

positives, ou son ordre de grandeur peut être approximativement connu, influençant ainsi le choix approprié de la distribution a priori.

Dans notre contexte spécifique de tarification des contrats de santé complémentaires collectifs, comme démontré dans le chapitre précédent, l'abondance des données statistiques disponibles facilite grandement le choix des distributions a priori. Ces données constituent une base solide pour dériver des distributions a priori à la fois informatives et réalistes. Elles permettent une intégration adéquate des tendances et particularités observées dans les modèles de sinistralité, tout en adhérant aux normes de prudence et de précision statistique. Cette richesse informationnelle est essentielle pour affiner nos modèles bayésiens, en veillant à ce qu'ils reflètent fidèlement les réalités du marché de l'assurance santé collective.

Le choix des distributions a priori transcende une simple procédure statistique, incarnant une démarche fondamentale qui intègre expertise actuarielle, connaissance théorique, et analyse empirique. Ce processus vise à produire des estimations fiables et véritablement représentatives des risques dans le domaine de l'assurance. Cette approche méthodique est cruciale pour assurer la pertinence et l'exactitude de nos modèles. Par la suite, nous explorerons les lois statistiques couramment utilisées, et nous verrons comment elles s'intègrent dans notre étude pour enrichir et affiner notre analyse.

- **Loi de Bernoulli** :

La loi de Bernoulli est une distribution de probabilité discrète simple, qui prend deux valeurs possibles, 1 (succès) et 0 (échec).

La probabilité d'une variable aléatoire X qui suit la loi de Bernoulli est définie comme :

$$p\left(X=x\right)=p^{x}\left(1-p\right)^{1-x} \quad ,x\in\{0;1\}$$

où x est la valeur de l'événement, prenant 1 en cas de succès et 0 en cas d'échec, et p est la probabilité du succès.

L'espérance de la loi de Bernoulli est p, et la variance de cette loi est $p\left(1-p\right)$. Cette distribution est souvent utilisée pour modéliser des événements binaires dans divers domaines, y compris en assurance pour représenter la survenance ou la non-survenance d'un sinistre.

- **Loi de Poisson** :

La loi de Poisson est une distribution de probabilité discrète essentielle en statistique, particulièrement pertinente dans le domaine de l'assurance pour modéliser le nombre d'événements (comme les sinistres) survenant dans un intervalle de temps fixe ou un espace spécifique, lorsque ces événements sont rares et indépendants les uns des autres.

La probabilité d'observer k événements est donnée par :

$$p\left(X=k\right)=\frac{\lambda^{k}e^{-\lambda}}{k!}$$

où λ est le taux moyen d'événements par intervalle, e est la base du logarithme naturel, et $k!$ est la factorielle de k.

L'espérance de la loi de Poisson est λ. Cela indique le nombre moyen d'événements attendus. La variance de cette loi est également λ, ce qui signifie que la variance est égale à la moyenne, une caractéristique unique de la loi de Poisson.

- **Loi Bêta** :

La loi Bêta est une distribution de probabilité continue qui est particulièrement utile pour modéliser des variables aléatoires dont les valeurs sont bornées entre 0 et 1, typiquement utilisée pour représenter une probabilité.

La densité de probabilité de la loi Bêta est donnée par :

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

où α et β sont des paramètres positifs qui contrôlent la forme de la distribution et $B(\alpha,\beta)$ est la fonction bêta, qui normalise la distribution. Lorsque $\alpha = \beta = 1$, la loi Bêta se réduit à une loi uniforme sur l'intervalle $[0,1]$.

Des valeurs plus élevées de α par rapport à β penchent la distribution vers 1, et inversement. L'espérance de la loi Bêta est $\frac{\alpha}{\alpha+\beta}$, et la variance de cette loi est $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Cette distribution est souvent choisie dans les modèles actuariels et financiers pour sa flexibilité et sa capacité à modéliser des proportions ou des probabilités.

- **Loi Gamma** :

La loi Gamma est une distribution de probabilité continue, souvent utilisée dans les analyses actuarielles et financières pour modéliser des variables dont les valeurs sont strictement positives, telles que les durées ou les montants de sinistres.

La densité de probabilité de la loi Gamma est définie par :

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1}\beta^{\alpha}e^{-\beta x}$$

où α est le paramètre de forme, β le paramètre d'intensité, et $\Gamma(\alpha)$ est la fonction gamma.

Le paramètre α détermine la forme de la distribution, tandis que β ajuste l'échelle et contrôle la dispersion de la distribution. L'espérance de la loi Gamma est $\frac{\alpha}{\beta}$, et la variance de cette loi est $\frac{\alpha}{\beta^2}$.

Cette distribution est particulièrement utile pour modéliser des données telles que les montants de sinistres ou les durées de processus, offrant une grande flexibilité pour s'adapter à diverses formes de données observées.

- **Loi normale et loi demi-normale** :

La loi normale, également connue sous le nom de distribution gaussienne, est l'une des distributions de probabilité les plus connues et les plus utilisées en statistique.

La densité de probabilité de la loi normal est donnée par :

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

où μ est la moyenne et σ est l'écart-type.

La loi demi-normale, une variante de la loi normale, est utilisée pour des variables qui ne prennent que des valeurs positives. Elle est définie par une distribution normale tronquée à zéro.

$$f(x;\sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} \qquad pour\ x \geq 0$$

L'espérance de la loi demi-normale est $\sigma\sqrt{\frac{2}{\pi}}$, et la variance de cette loi est $\sigma^2\left(1-\frac{2}{\pi}\right)$. Cette distribution est utile pour modéliser des situations où la distribution d'une variable est centrée autour de zéro mais ne prend pas de valeurs négatives, tout en étant idéale pour représenter l'écart-type ou la variance d'une variable aléatoire.

Modèle de mélange

Dans un modèle de mélange, on reconnaît que les données peuvent provenir de différentes sous-populations, chacune avec sa propre distribution statistique. L'observation que nous mesurons est donc considérée comme étant tirée de l'une de ces distributions sous-jacentes, et le modèle de mélange cherche à identifier non seulement les paramètres de ces distributions, mais aussi la proportion dans laquelle chaque distribution contribue à l'ensemble des données.

Mathématiquement, un modèle de mélange peut être exprimé comme une somme pondérée des distributions de probabilité de chaque sous-population. Pour une variable aléatoire Y, le modèle de mélange peut être écrit comme :

$$p(Y|\Theta) = \sum_{k=1}^{K} \pi_k \cdot f_k(Y|\theta_k)$$

où K est le nombre de distributions dans le mélange, π_k est la probabilité qu'une observation provienne de la k-ème sous-population, f_k est la densité de probabilité de la k-ème distribution, et θ_k sont les paramètres de cette distribution.

En inférence bayésienne, on utilise des distributions a priori pour les paramètres θ_k et pour les probabilités de mélange π_k, et on applique les MCMC pour obtenir des échantillons de la distribution a posteriori, ce qui permet d'inférer la composition des sous-populations au sein des données observées.

Cas particulier : modèle à inflation de zéros

Le modèle à inflation de zéros est une application particulière de modèle de mélange, adaptée aux données de comptage avec un excès de zéros. Ils combinent une distribution de comptage, comme la loi de Poisson ou la loi binomiale négative, avec une distribution de Bernoulli pour modéliser la probabilité qu'une observation soit un zéro, non dû au hasard mais à des facteurs spécifiques. Le modèle postule ainsi deux processus : un processus de génération de zéros, et un processus de comptage pour les valeurs non nulles.

En contexte bayésien, on peut construire des distributions a priori pour la probabilité d'inflation de zéros et pour les paramètres de la distribution de comptage. Les techniques MCMC sont utilisées pour surmonter les défis de calcul de la distribution a posteriori et fournissent une méthode robuste pour estimer les paramètres du modèle et prédire de nouvelles observations. Ce modèle est bien utile dans les domaines où les zéros ne sont pas simplement l'absence de comptage, mais portent également une signification, comme l'absence de sinistres dans les données d'assurance.

Modèle hiérarchique bayésien

Le modèle hiérarchique bayésien est une forme de modélisation statistique écrite en plusieurs niveaux qui estime les paramètres de la distribution a posteriori en utilisant la méthode bayésienne. Cette approche intègre divers sous-modèles qui se combinent pour constituer une architecture globale multicouche. Le théorème de Bayes joue un rôle essentiel en fusionnant ces sous-modèles avec les données collectées, tout en gérant l'incertitude inhérente à chaque niveau. Cette fusion produit la distribution a posteriori, qui est essentiellement une version raffinée de l'estimation probabiliste initiale, mise à jour en continu à mesure que de nouvelles données viennent affiner ou modifier la compréhension de la distribution a priori.

La modélisation hiérarchique est utilisée lorsque des informations sont disponibles sur plusieurs niveaux différents d'unités d'observation. La structure des données pour la modélisation hiérarchique conserve une structure de données imbriquée. La forme hiérarchique d'analyse et d'organisation aide à comprendre les problèmes à plusieurs paramètres et joue également un rôle important dans le développement de stratégies de calcul.

Pour illustrer concrètement le fonctionnement d'un modèle hiérarchique bayésien, considérons un cadre simplifié où y_i représente une observation, et θ_i est le paramètre qui régit le processus de génération de données pour y_i. Ces paramètres $\theta_1, \theta_2, ..., \theta_j$ sont supposés être générés de manière interchangeable à partir d'une population commune, avec une distribution gouvernée par un hyperparamètre ϕ. La modélisation hiérarchique bayésienne se déroule en trois étapes :

- Étape 1 : $y_i|\theta_i, \phi \sim P(y_i|\theta_i, \phi)$ - C'est la vraisemblance de l'observation y_i étant donnée les paramètres θ_i et l'hyperparamètre ϕ.
- Étape 2 : $\theta_i|\phi \sim P(\theta_i|\phi)$ - Cela représente la distribution a priori de θ_i conditionnelle à ϕ, reflétant les croyances avant de voir les données.
- Étape 3 : $\phi \sim P(\phi)$ - C'est la distribution a priori de l'hyperparamètre ϕ.

La vraisemblance est donc dépendante de ϕ uniquement à travers θ_i, et la distribution a priori peut être décomposée en $P(\theta_i, \phi) = P(\theta_i|\phi) P(\phi)$ en utilisant la définition de la probabilité conditionnelle.

Ainsi, la distribution a posteriori $P(\phi, \theta_i|y)$ est proportionnelle à $P(y_i|\theta_i) P(\theta_i|\phi) P(\phi)$, qui est calculée en utilisant le théorème de Bayes.

Dans ce cadre, les hyperparamètres jouent un rôle crucial car ils permettent de lier les niveaux hiérarchiques du modèle, apportant une cohérence globale et permettant de "partager l'information" entre les différentes unités d'observation à travers la structure hiérarchique du modèle.

3.3 Application pratique sur la tarification a posteriori

Dans cette section, nous allons explorer l'application pratique de l'approche bayésienne pour la tarification a posteriori dans le secteur de l'assurance santé collective. Avec les fondements de la tarification et les techniques des modèles bayésiens établis dans les sections précédentes, nous sommes désormais en mesure d'estimer et d'ajuster les primes pures en utilisant les données historiques.

Bien que la tarification a posteriori puisse être calculée à différents niveaux, comme pour un bénéficiaire individuel, un groupe d'entreprises, ou même l'ensemble des entreprises d'une formule

de produit santé, notre attention se portera spécifiquement sur le niveau d'entreprise. Cette approche ciblée est pertinente puisque, dans le contexte de l'assurance santé collective, le calcul de la prime pure a posteriori pour un individu seul ne serait pas pertinent. De même, bien que le calcul à l'échelle de toutes les entreprises d'une formule puisse être utile pour examiner les problèmes de rentabilité, ce n'est pas l'objectif de notre étude. De plus, les différences de primes entre les entreprises au sein d'une même formule, dues à des facteurs tels que l'âge moyen, la zone géographique ou l'abattement commercial, rendent difficile l'interprétation des primes pures a posteriori pour chaque entreprise.

L'accent sur la maille entreprise s'avère donc être le plus approprié. En général, au sein d'une même entreprise, les montants de la prime sont identiques pour tous les employés, et les primes pures a priori ont déjà été ajustées en tenant compte de facteurs tels que l'âge moyen, le secteur d'activité et la localisation géographique. En comparant les primes pures a posteriori et a priori, nous pourrons mieux comprendre l'hétérogénéité des profils de risque, surtout pour des entreprises partageant des caractéristiques similaires.

Dans les pages suivantes, nous illustrerons concrètement l'utilisation des modèles bayésiens pour calculer la prime pure a posteriori pour une entreprise donnée. Étant donné la complexité et la diversité des regroupements d'actes, qui peuvent être au nombre d'une quarantaine, nous nous concentrerons spécifiquement sur trois regroupements d'actes distincts pour démontrer de manière concrète l'utilisation de ces modèles en tarification a posteriori.

Regroupement d'actes - Lentilles de contact

Dans cette partie, nous allons nous concentrer sur le calcul de la prime pure a posteriori pour une entreprise sélectionnée aléatoirement. Notre démarche consistera à estimer séparément la fréquence annuelle et le coût associé, spécifiquement pour le regroupement d'actes lié aux lentilles de contact. Nous illustrerons le processus complet de cette estimation : depuis la sélection des distributions a priori adaptées, en passant par la définition de la vraisemblance, jusqu'à l'analyse de convergence et la simulation des données via la distribution a posteriori. Cette approche détaillée et méthodique est essentielle pour assurer une compréhension approfondie et précise du mécanisme de tarification a posteriori bayésienne en actuariat.

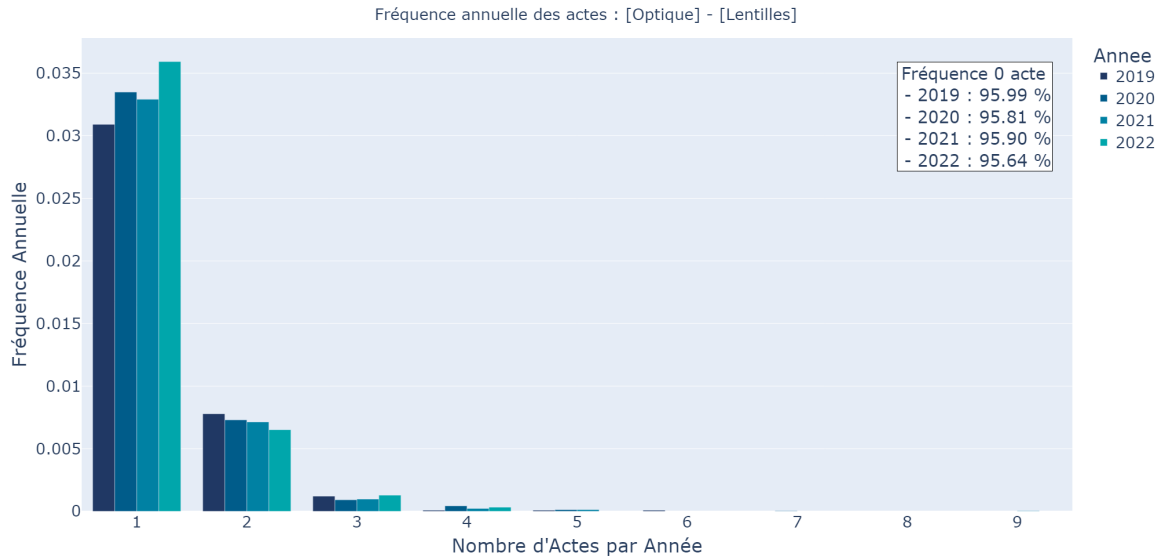
- **Estimation bayésienne de la fréquence annuelle** :

Pour estimer la fréquence annuelle des actes concernant les lentilles de contact, il est primordial de commencer par le choix d'une distribution a priori. Cette distribution incarne nos connaissances préalables et nos croyances avant l'analyse des données. La sélection du type de cette distribution, ainsi que de ses paramètres, est généralement guidée par des statistiques globales, provenant d'études internes réalisées sur l'ensemble des assurés.

Cependant, dans le but de simplifier l'étude et en raison de la spécificité de notre périmètre, nous optons pour une approche plus ciblée. Plutôt que de nous appuyer sur des statistiques globales, qui peuvent ne pas être représentatives pour un segment spécifique, nous utilisons des données propres à notre périmètre. Ces données, plus représentatives de notre segment d'étude, permettent d'affiner notre modèle bayésien, assurant ainsi des estimations plus précises et pertinentes pour la fréquence des actes liés aux lentilles de contact.

Le graphique 3.1 ci-dessous illustre les fréquences annuelles des actes relatifs aux lentilles de contact de 2019 à 2022. Il montre que, dans notre périmètre d'étude, seulement environ 4% de la population effectue des actes liés aux lentilles de contact. Les données concernant la fréquence de 0 acte sont

FIGURE 3.1 – Répartition des fréquences des lentilles de contact



exclues des barres pour plus de clarté et sont détaillées dans l'encart en haut à droite.

Parmi la population concernée, il est courant de voir des individus engager ces actes une ou deux fois par an. Bien qu'une loi de Poisson puisse être envisagée pour estimer cette distribution, elle pourrait sous-estimer les occurrences plus fréquentes. En tant qu'alternative, une combinaison de distributions pourrait être plus adaptée. Néanmoins, dans une démarche initiale, nous opterons pour une loi de Poisson avec un paramètre lambda considéré comme une variable aléatoire, ce qui souligne la flexibilité des modèles bayésiens.

Nous allons maintenant porter notre attention sur une entreprise sélectionnée aléatoirement pour notre étude de cas. Comme le révèle le tableau 3.1 suivant, nous avons rassemblé les données concernant le nombre de bénéficiaires et la fréquence de leurs achats de lentilles de contact de l'année 2019 à 2022.

TABLE 3.1 – Répartition de consommation de lentilles par bénéficiaire - Entreprise Exemple

Année d'observation	**Nombre d'actes**				
	0	**1**	**2**	**3**	**>3**
2019	174	24	0	0	0
2020	178	30	0	0	0
2021	191	17	0	1	0
2022	193	18	0	0	0

Il est à noter que la majorité des bénéficiaires ne réalise aucun achat de lentilles de contact au cours d'une année donnée. Pour ceux qui en achètent, les proportions atteignant au moins une consommation annuelle sont de 12,1% en 2019, 14,4% en 2020, 8,6% en 2021, et 8,5% en 2022. Ces chiffres dépassent la moyenne observée dans notre périmètre d'étude, qui est d'environ 4%.

Basés sur ces observations et en nous appuyant sur notre connaissance générale des entreprises, nous élaborerons un modèle bayésien. Ce modèle cherchera à affiner la caractérisation et la compréhension des schémas de consommation propres à l'entreprise étudiée, en trouvant un équilibre entre les

connaissances générales issues des statistiques globales et les spécificités propres à cette entreprise.

Adoptons une structure élémentaire illustrée par le schéma 3.2 ci-après : la fréquence annuelle d'achats de lentilles de contact est modélisée par une loi de Poisson caractérisée par un paramètre lambda. Ce dernier est lui-même déterminé par une loi Bêta. Sans préconception sur la distribution de la fréquence d'achats, nous optons initialement pour une loi Bêta(1,1), équivalente à une distribution uniforme entre 0 et 1. Le modèle est ensuite alimenté par nos données historiques d'achat de lentilles de contact, qui comprennent 826 instances, révélant que la majorité de ces transactions correspondent à des cas où aucun achat n'a été effectué.

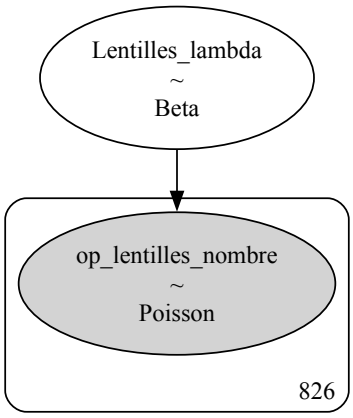
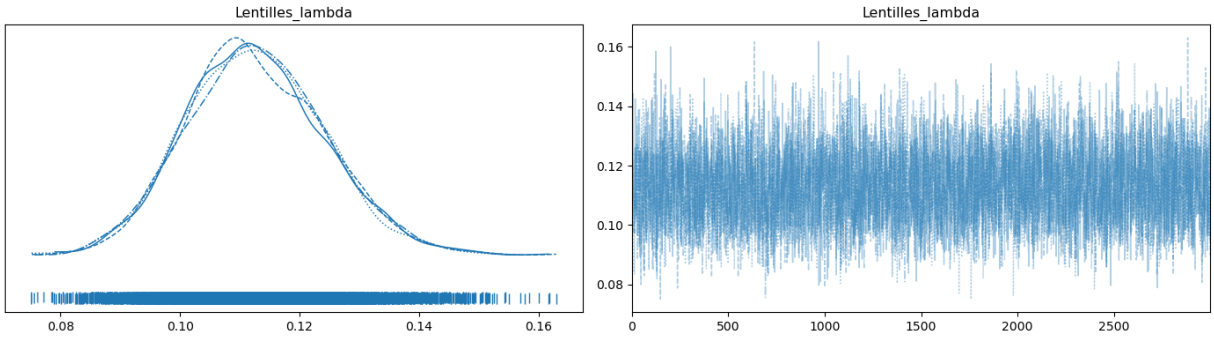


FIGURE 3.2 – Schéma du premier modèle des fréquences des lentilles de contact

Les graphiques 3.3 présentés ci-dessous révèlent des détails clés de notre modèle bayésien élémentaire, qui évalue la fréquence d'achat de lentilles de contact. Le graphique de gauche montre la distribution a posteriori du paramètre lambda. En l'absence de connaissances préétablies, cette distribution capte adéquatement une moyenne des données historiques sur les quatre dernières années, s'étendant de 8% à 15%.

FIGURE 3.3 – Convergence et distribution de Lambda - Distribution a priori non-informative

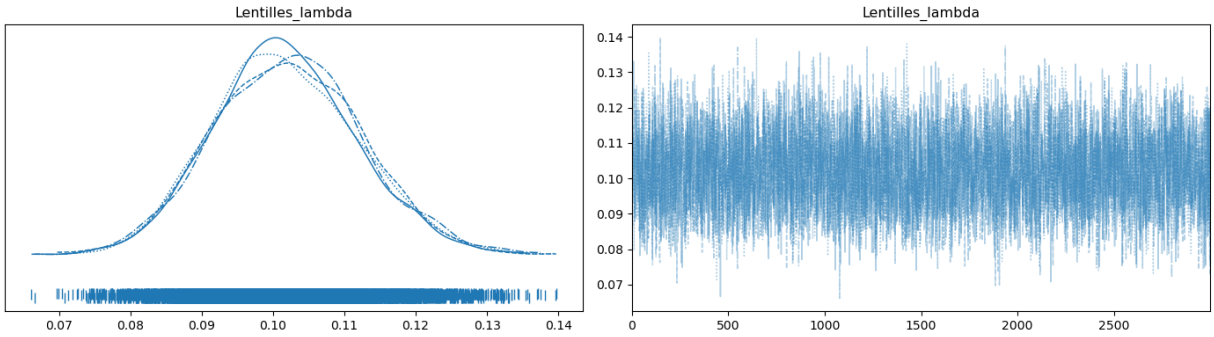


Le graphique de droite détaille les échantillons issus des chaînes de Markov, disposés en séquence. Cette représentation séquentielle est cruciale pour vérifier la convergence de l'algorithme d'échantillonnage. Une convergence correcte est indispensable pour assurer que les échantillons soient représentatifs de la distribution a posteriori. La densité des échantillons et la cohérence dans la variation suggèrent que la chaîne a atteint une stabilité suffisante et que la méthode d'échantillonnage a bien convergé, ce

qui renforce la crédibilité des estimations du modèle pour le paramètre lambda.

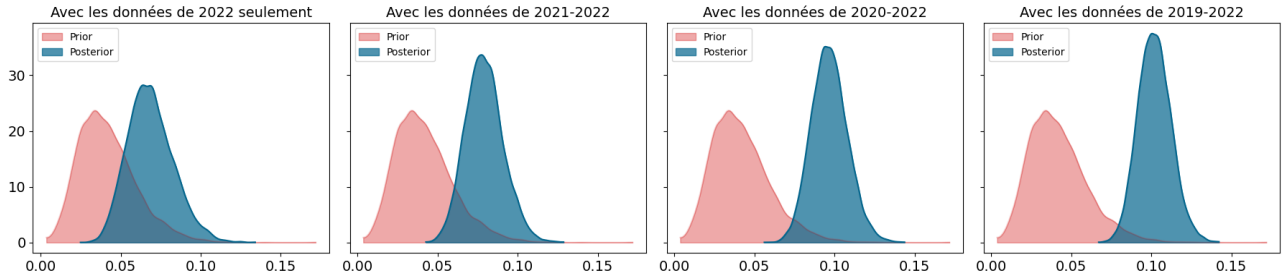
Nous allons désormais incorporer une distribution a priori plus informative, reflétant notre compréhension préalable de la fréquence d'achat de lentilles, avant de prendre en compte les données historiques de l'entreprise concernée. La sélection de la distribution a priori peut varier selon les perspectives individuelles, et elle demeure intrinsèquement subjective. Dans notre cas, nous ne nous attarderons pas à déterminer la distribution a priori la plus adéquate. Nous postulons que la distribution du paramètre lambda est régie par une loi Bêta(5,115), basée sur les statistiques globales de l'industrie. En intégrant les données historiques de l'entreprise spécifique à notre étude, nous recalibrerons le modèle. Les résultats mis à jour concernant la convergence et la distribution de lambda sont illustrés dans le graphique 3.4 qui suit.

FIGURE 3.4 – Convergence et distribution de Lambda - Distribution a priori informative



Il est observé que la distribution de lambda s'est décalée par rapport à la moyenne historique centrée à 11%. La moyenne de la distribution a posteriori s'est en effet orientée vers la gauche, s'établissant désormais à 10%, ce qui est légèrement en dessous de la moyenne observée pour l'entreprise en question. Quant au graphique de droite, il confirme que la chaîne de Markov a atteint une stabilité adéquate, signifiant que la méthode d'échantillonnage a correctement convergé.

FIGURE 3.5 – Évolution de la distribution de lambda avec l'intégration des données



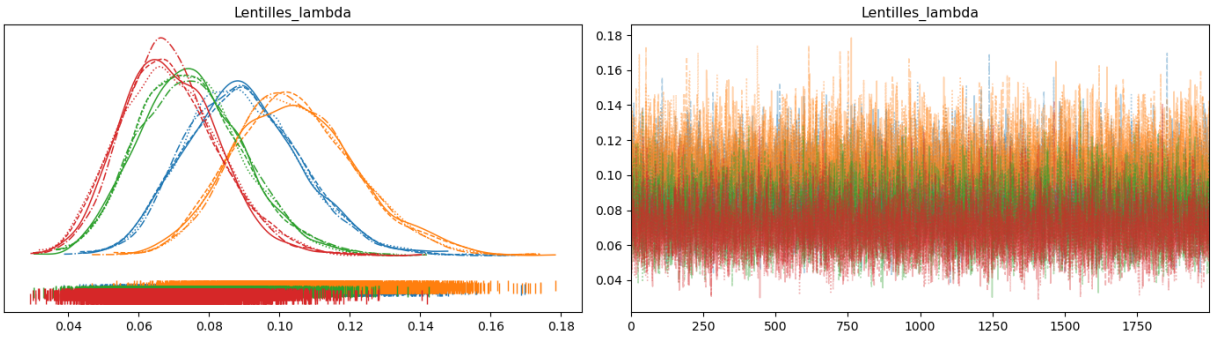
Le graphique 3.5 ci-dessus montre comment l'intégration progressive des données historiques influence la distribution a posteriori du paramètre lambda, en la comparant à la distribution a priori. Cette évolution met en évidence l'impact des nouvelles données sur l'ajustement des estimations du modèle. Sur l'extrême gauche, le graphique révèle l'influence de l'utilisation exclusive des données de 2022. On constate que la distribution a posteriori reste relativement proche de la distribution a priori, traduisant ainsi une plus grande incertitude par rapport aux autres graphiques.

À mesure que l'on inclut davantage de données historiques, de 2021 à 2019, l'incertitude décroît et la

distribution a posteriori se rapproche davantage du comportement spécifique de l'entreprise, s'écartant de la distribution a priori, surtout lorsque celle-ci diffère significativement de la moyenne générale due à une fréquence d'achats nettement supérieure. Cela démontre un principe fondamental de la tarification bayésienne a posteriori, où l'ajout de données vient affiner et ajuster nos estimations en fonction des tendances spécifiques observées.

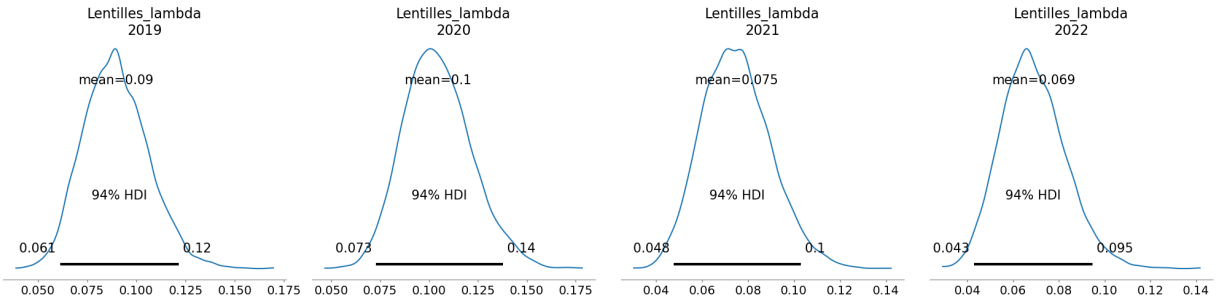
Il est essentiel de reconnaître que, même au sein d'une même entreprise, les comportements peuvent varier d'une année sur l'autre. Ces variations peuvent être dues à des tendances générales, telles qu'une surconsommation entraînée par une hausse annuelle de la fréquence des achats. Elles peuvent également résulter de circonstances ou d'événements ponctuels, à l'instar de la pandémie de Covid-19. Les changements observés peuvent aussi être spécifiques à un secteur particulier, à une région géographique donnée, ou affecter uniquement l'entreprise en question. Nous allons maintenant examiner s'il existe une variation notable dans la fréquence de consommation de lentilles de contact pour l'entreprise étudiée.

FIGURE 3.6 – Convergence et distribution de Lambda - Années 2019 à 2022



Les graphiques 3.6 ci-dessus illustrent les distributions a posteriori du paramètre lambda pour les années 2019 à 2022, révélant de légères différences entre les courbes annuelles. Pour une entreprise de taille moyenne, de telles variations sont généralement attendues. Toutefois, pour affiner la tarification future, il peut être judicieux de donner un poids plus conséquent aux données plus récentes par rapport aux années antérieures, reflétant une évolution potentielle ou des changements récents dans les tendances de consommation. Le graphique de droite démontre que les chaînes de Markov ont abouti à une stabilité suffisante, indiquant une convergence réussie de la méthode d'échantillonnage.

FIGURE 3.7 – Distribution a posteriori de Lambda - Années 2019 à 2022



Les graphiques 3.7 affichent la distribution a posteriori du paramètre lambda de l'année 2019 à 2022. Chaque graphique montre la moyenne de la distribution a posteriori pour chaque année respective, indiquant la fréquence moyenne d'achat estimée par le modèle bayésien. En 2019 et 2020, nous observons

des moyennes plus élevées (9% et 10% respectivement), suggérant une fréquence d'achat supérieure pendant ces années. En 2021 et 2022, la moyenne diminue (7,5% et 6,9% respectivement), indiquant une baisse possible dans la fréquence d'achat.

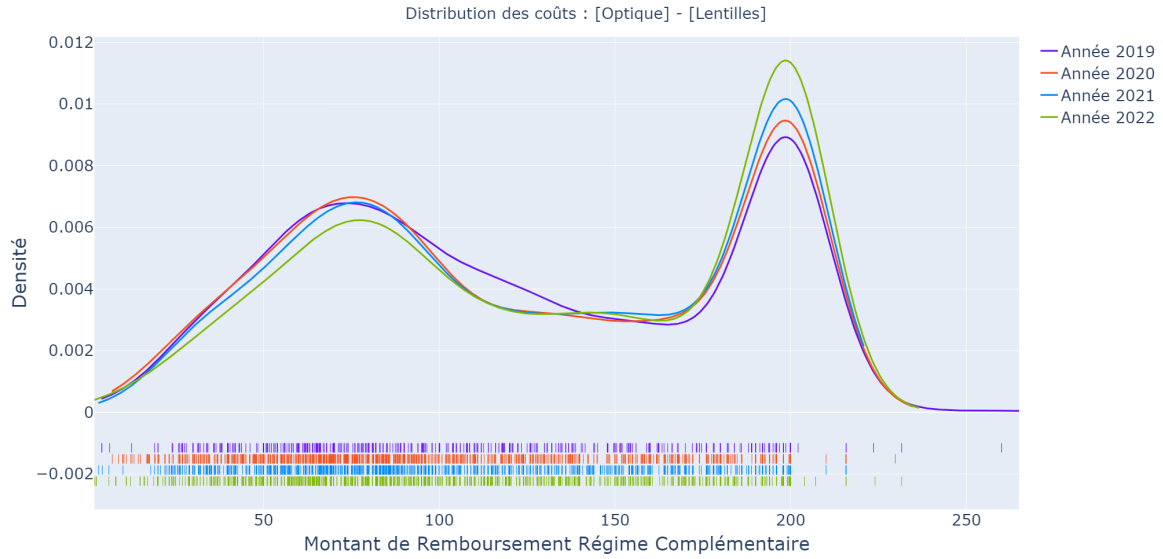
Le terme "HDI" fait référence à l'intervalle de densité le plus élevé à 94%, c'est-à-dire l'intervalle le plus court possible contenant 94% de la distribution a posteriori. Cet intervalle est utilisé pour représenter l'incertitude de l'estimation de lambda, plus cet intervalle est étroit, plus la prédiction est précise. L'HDI est une alternative bayésienne à l'intervalle de confiance classique, et il fournit une plage crédible pour la valeur du paramètre lambda en se basant sur les données observées.

À l'aide d'un modèle bayésien, nous avons dérivé la distribution a posteriori pour le paramètre lambda. Cette distribution nous permet de générer des simulations pour ce paramètre et, par extension, de simuler la fréquence d'achat de lentilles de contact pour l'entreprise étudiée. L'étape suivante consistera à simuler les coûts y afférents. En multipliant les nombres d'achats annuels estimés par les coûts associés, nous obtiendrons une série de simulations pour le coût de la garantie des lentilles. En calculant la moyenne de ces simulations, nous serons en mesure de fixer la prime pure pour cette garantie, intégrant à la fois les connaissances du périmètre global et les spécificités de l'entreprise concernée.

- **Estimation bayésienne du coût associé** :

Dans cette partie, nous abordons la question de l'estimation bayésienne des coûts associés aux lentilles de contact pour l'entreprise en question. Nous débuterons en examinant la distribution globale des coûts des lentilles, prenant en compte l'ensemble du périmètre concerné. Cela nous fournira un aperçu des dépenses typiques et servira de point de départ pour estimer les coûts propres à chaque entreprise. En intégrant ces informations dans notre modèle bayésien, nous serons à même de produire une estimation des coûts ajustée et personnalisée qui reflète non seulement les tendances du périmètre mais aussi les particularités de l'entreprise étudiée.

FIGURE 3.8 – Répartition des remboursements des lentilles de contact



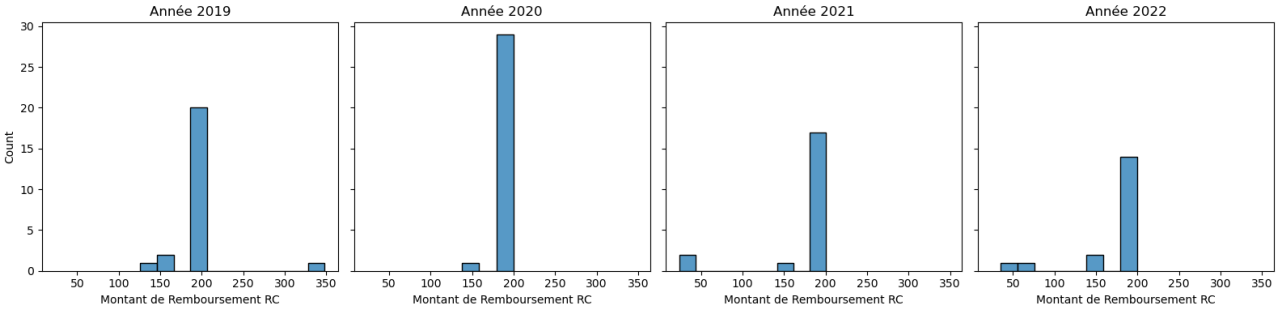
Le graphique 3.8 illustre la distribution des montants de remboursement du régime complémentaire pour les lentilles de contact sur quatre années consécutives, de 2019 à 2022. Comme cela a été

discuté dans le chapitre 2 concernant les soins dentaires non remboursés, nous observons ici aussi une distribution tronquée à 200 euros, indiquant que les remboursements sont plafonnés à ce montant avec quelques valeurs qui s'écartent de cette limite, représentant probablement des anomalies ou des cas exceptionnels.

Les courbes de densité pour chaque année sont très proches les unes des autres, suggérant une grande similitude dans les montants de remboursement d'année en année. Cette proximité indique qu'il n'y a pas eu de changement significatif dans les comportements de consommation ou dans les politiques de remboursement des lentilles de contact au cours de cette période. Les légères fluctuations observées sur le graphique dénotent une légère tendance à l'augmentation des remboursements au plafond sur la période des quatre dernières années, bien que cette tendance soit mineure.

Maintenant, en nous concentrant sur l'entreprise étudiée, le graphique 3.9 ci-dessous indique une tendance intéressante. Bien que les données proviennent d'une entreprise de taille moyenne, où il est courant de ne pas disposer d'un volume important d'informations, nous observons un modèle de remboursement remarquable. Lors des achats de lentilles de contact, le montant de remboursement atteint fréquemment 200 euros, ce qui correspond au plafond de la garantie. Ce schéma est différent de celui observé dans l'ensemble des entreprises du périmètre, soulignant un comportement spécifique où les montants de remboursement tendent à être maximisés. Dans cet exemple, nous ne pouvons pas déterminer si ce comportement est propre à l'entreprise en question ou s'il est lié à une zone géographique en particulier. Cependant, un modèle bayésien hiérarchique pourrait être utile pour identifier l'effet de la zone géographique, de la démographie ou de l'entreprise.

FIGURE 3.9 – Répartition des remboursements des lentilles - Entreprise Exemple

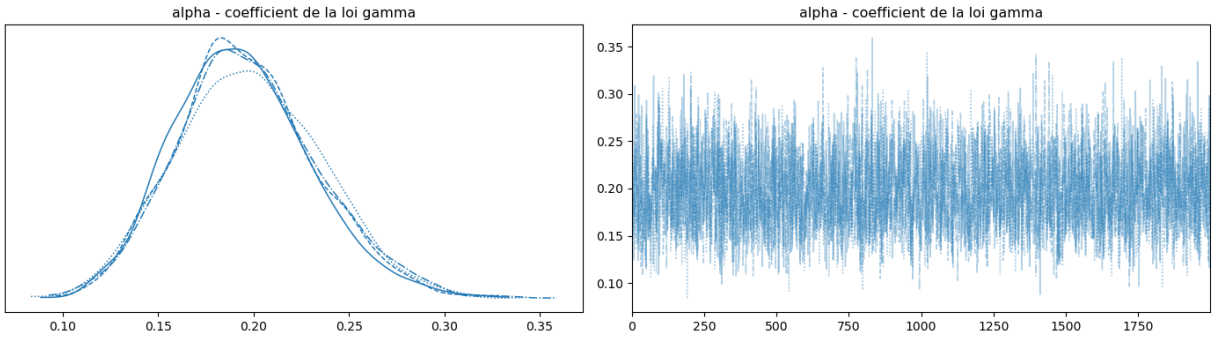


Pour représenter la distribution des montants de remboursement des lentilles de contact, nous devons tenir compte du plafond de garantie fixé à 200 euros. Une loi statistique classique peut ne pas être adéquate pour capturer cette spécificité. Dans cet exemple, nous allons donc appliquer la méthode du modèle de mélange, tel que décrit dans ce chapitre.

La distribution des montants est modelée par la combinaison d'une loi gamma, appropriée pour les distributions ne prenant que des valeurs positives, et d'une constante à 200 euros qui correspond au plafond de garantie. La loi gamma est pondérée par un coefficient α, tandis que la probabilité d'atteindre la constante de 200 euros est représentée par $1 - \alpha$. Dans une approche bayésienne, α est lui-même une variable aléatoire dont la distribution a priori doit être définie. Compte tenu de sa nature variant de 0 à 1, une distribution Bêta est adéquate pour modéliser cette distribution a priori.

Le graphique 3.10 de gauche montre la distribution a posteriori du coefficient α, qui détermine le poids de la loi gamma dans le modèle de mélange pour les remboursements de lentilles de contact. Il semble que près de 20% des observations soient mieux décrites par la loi gamma, tandis que les 80%

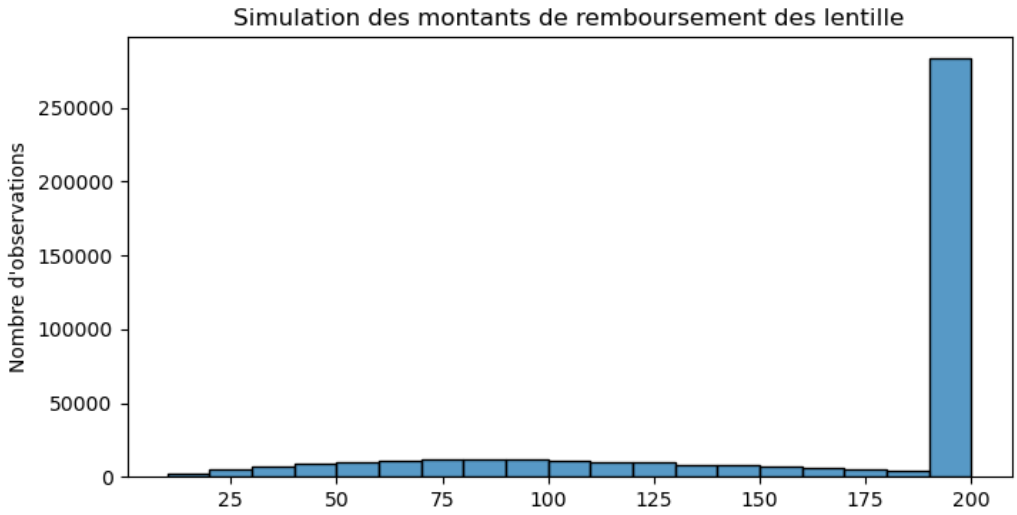
FIGURE 3.10 – Convergence et distribution du coefficient alpha



restants sont alignés sur le montant de 200 euros, qui est le plafond de la garantie pour les lentilles. Ce ratio est inférieur à celui observé pour l'ensemble du secteur, mais il est plus élevé que celui spécifique à l'entreprise analysée, offrant ainsi un équilibre entre les expériences générales et les données propres à une entreprise donnée.

Le graphique de droite, quant à lui, démontre la convergence des chaînes de Markov, indiquant que la méthode d'échantillonnage a atteint une stabilité et que les échantillons sont représentatifs de la distribution a posteriori. Cela confirme la fiabilité des estimations du coefficient α dérivées du modèle bayésien.

FIGURE 3.11 – Simulation des montants de remboursement de lentilles



Le graphique 3.11 affiche la distribution simulée des montants de remboursement pour les lentilles de contact par le régime complémentaire. Comme le démontre la graphique, un pic prononcé apparaît à 200 euros, ce qui est cohérent avec le fait de plafonnement des remboursements. Les simulations révèlent également des montants de remboursement s'échelonnant de 1 à 199 euros, indiquant que même si ces montants sont peu représentés dans les données historiques de l'entreprise analysée, ils émergent dans les simulations grâce à l'incorporation de la distribution a priori. Cela indique que notre modèle produit une estimation équilibrée qui n'est pas faussée par une sur-représentation des remboursements de l'entreprise spécifique ; il capture adéquatement à la fois l'influence générale du marché et les particularités individuelles de l'entreprise.

À travers les simulations effectuées sur le nombre d'achats de lentilles de contact et les coûts y afférents, nous sommes en mesure de générer des estimations des montants annuels de remboursement pour ces lentilles, par bénéficiaire. Cette estimation se base sur le calcul du produit entre le nombre d'achats et le coût associé. En évaluant la moyenne de ces montants simulés, nous déterminons la prime pure a posteriori pour la garantie des lentilles, comme illustré dans le tableau 3.2 ci-dessous. Cette prime pure représente un équilibre entre une tarification globale, fondée sur des distributions a priori, et une tarification personnalisée qui intègre les données historiques de l'entreprise, permettant ainsi une approche plus ajustée et représentative des besoins réels des bénéficiaires.

TABLE 3.2 – Prime pure a posteriori par bénéficiaire - Lentilles de contact

Regroupement des actes	**Prime pure a posteriori**
Lentilles de contact	16,11 € par an par bénéficiaire

Regroupement d'actes - Kinésithérapie

Dans cette section, nous explorerons un autre cas d'étude : la kinésithérapie. Comme discuté au chapitre deux, le regroupement d'actes en kinésithérapie se distingue par sa faible occurrence (environ 10%), mais pour les bénéficiaires concernés, le volume annuel de soins peut être considérable.

Notre approche, similaire à celle utilisée pour les lentilles, consistera à évaluer de manière distincte la fréquence annuelle et le coût correspondant pour ce regroupement d'actes en kinésithérapie. Nous décrirons en détail le processus d'estimation, allant de la sélection de distributions a priori appropriées à la définition de la vraisemblance, et finalement à l'analyse de la convergence et la simulation des données en utilisant la distribution a posteriori. Cette méthode rigoureuse et approfondie est cruciale pour garantir une compréhension exhaustive et précise des principes de tarification bayésienne a posteriori dans le domaine actuariel.

- **Estimation bayésienne de la fréquence annuelle** :

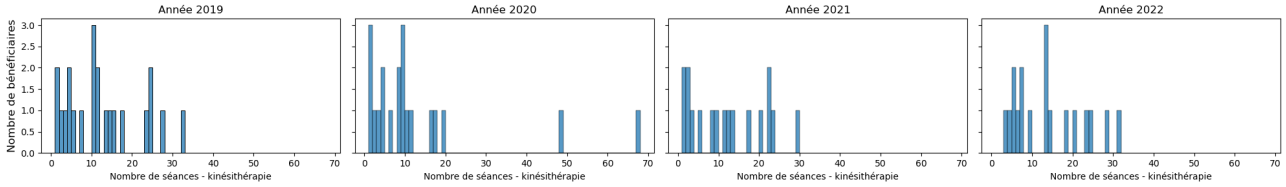
Dans le chapitre précédent, nous avons introduit la distribution des séances de kinésithérapie. Il est essentiel de souligner que seulement 10% de la population totale des entreprises du périmètre de l'étude recourent à la kinésithérapie, et parmi eux, certains peuvent nécessiter plus de 160 séances annuelles.

Cette distribution unique ne se prête pas à une modélisation par des lois statistiques simples. Nous proposons donc l'emploi d'un modèle de mélange, en particulier un modèle à inflation de zéros, pour mieux représenter ces données.

Notre objectif est de fusionner une loi discrète avec une surabondance de zéros. Après une analyse approfondie des données issues de l'ensemble des entreprises, nous avons choisi d'appliquer une loi géométrique, appropriée pour modéliser une distribution discrète où la probabilité diminue avec l'augmentation du nombre de séances. Pour la loi géométrique, le coefficient de pondération crucial est α, pour lequel nous avons choisi d'adopter une distribution a priori uniforme située entre 0 et 0.2. En outre, il est nécessaire d'attribuer également une distribution a priori au paramètre p de la loi géométrique, qui étant une probabilité, sera modélisée par une loi Bêta.

Cette approche bayésienne permet une estimation plus précise et adaptée de la fréquence annuelle des séances de kinésithérapie, tenant compte de la spécificité et de la variabilité des données dans notre contexte actuariel.

FIGURE 3.12 – Répartition des séances kinésithérapie - Entreprise Exemple



Le graphique 3.12 illustre la distribution des séances de kinésithérapie pour l'entreprise étudiée, réparties sur quatre années consécutives, de 2019 à 2022. En observant les barres, il est clair que la majorité des bénéficiaires nécessitent un nombre de séances allant de 1 à 30 par an. Cependant, il est également évident qu'il existe une minorité de cas où le nombre de séances annuelles s'élève entre 40 et 70, suggérant des besoins de soins plus intensifs pour ces individus.

À travers les années, il semble y avoir une certaine cohérence dans la répartition des séances, avec une concentration notable de cas dans les gammes les plus basses de séances. Les pics sporadiques au-delà de 40 séances pourraient indiquer des cas isolés nécessitant des soins prolongés ou intensifs.

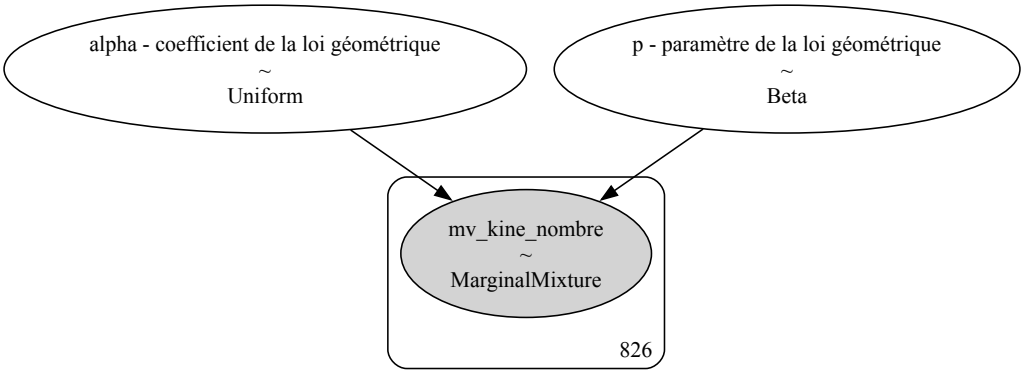
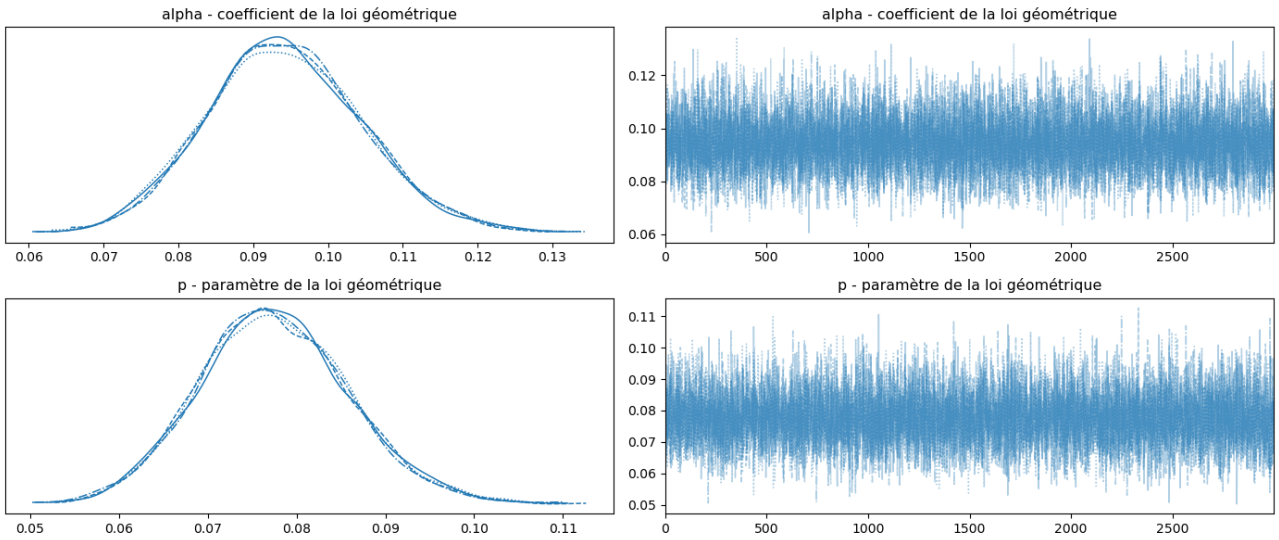


FIGURE 3.13 – Schéma du modèle des fréquences des séances de kinésithérapie

La distribution observée est en accord avec les tendances identifiées précédemment au niveau de l'ensemble des entreprises, ce qui renforce la pertinence de notre modèle initial pour cette analyse spécifique. Les données de l'entreprise en question semblent donc bien se prêter à l'application de notre approche de modélisation bayésienne. Le graphique 3.13 ci-dessus présente de manière schématique la structure du modèle que nous envisageons, offrant un aperçu visuel de la méthode et des paramètres impliqués dans l'estimation de la fréquence annuelle des séances de kinésithérapie. Cette représentation schématique servira de guide pour la mise en œuvre détaillée de notre modèle bayésien et pour l'interprétation des résultats qui en découlent.

FIGURE 3.14 – Convergence et distribution des paramètres

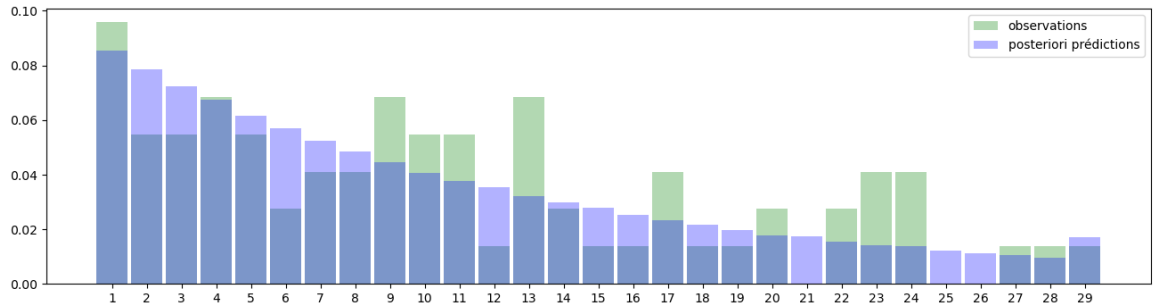


Les graphiques 3.14 ci-dessus montrent les distributions de densité de probabilité et les traces de convergence pour le coefficient α et le paramètre p de la loi géométrique. La distribution a posteriori du coefficient α se concentre autour de 9,5%, ce qui coïncide avec la proportion observée de patients nécessitant des séances de kinésithérapie au moins une fois par an. Cette convergence autour d'une valeur spécifique suggère que le modèle a correctement appris à partir des données et reflète le comportement réel des bénéficiaires.

Concernant le paramètre p, une valeur faible implique un nombre élevé de séances de kinésithérapie, comme on peut l'observer dans la distribution a posteriori. Cela correspond à des situations où certains patients requièrent un nombre significatif de séances, conformément aux données observées.

Les graphiques de droite illustrent la convergence des chaînes de Markov pour les deux paramètres. Ces traces, qui semblent atteindre un plateau sans tendance claire ou sans dérive, indiquent que les chaînes sont bien mélangées et que la méthode d'échantillonnage par chaînes de Markov a atteint une distribution stable. C'est une indication forte que les échantillons générés sont représentatifs de la distribution a posteriori et que les paramètres du modèle peuvent être considérés comme fiables pour la simulation et l'inférence statistique dans le cadre de l'estimation bayésienne de la fréquence annuelle de séances de kinésithérapie.

FIGURE 3.15 – Simulation de séance kinésithérapie - Entreprise Exemple



Le graphique 3.15 ci-dessus présente une superposition des données observées et des prédictions a posteriori pour le nombre de séances de kinésithérapie au sein de l'entreprise étudiée qui est une entreprise de taille moyenne. Les barres vertes représentent les observations réelles, tandis que les barres bleues indiquent les prédictions simulées à partir du modèle.

On constate que le modèle produit une simulation qui capture assez fidèlement la distribution réelle des nombres de séances, avec une tendance générale à la décroissance du nombre de bénéficiaires à mesure que le nombre de séances augmente. Cette adéquation suggère que le modèle est bien calibré et peut reproduire les tendances observées dans les données réelles de l'entreprise.

Il est à noter que certains écarts, sous forme de "trous" ou de "pics", apparaissent entre les observations et les simulations. Cela peut être attribué à la variabilité inhérente à une population d'entreprise de taille moyenne et ne diminue en rien la validité générale du modèle. Globalement, la correspondance entre les prédictions du modèle et les données observées témoigne de l'efficacité du modèle bayésien pour estimer le nombre de séances de kinésithérapie nécessaires.

- **Estimation bayésienne du coût associé** :

Nous nous tournons désormais vers la distribution des coûts associés aux séances de kinésithérapie. Afin de poser une base solide pour notre analyse, nous débutons par l'examen des données de remboursement de l'ensemble des entreprises du périmètre. Cette étape préliminaire nous permettra de comprendre la répartition globale des dépenses de kinésithérapie et de dégager des tendances ou des modèles pertinents qui pourraient influencer le coût total de ces services.

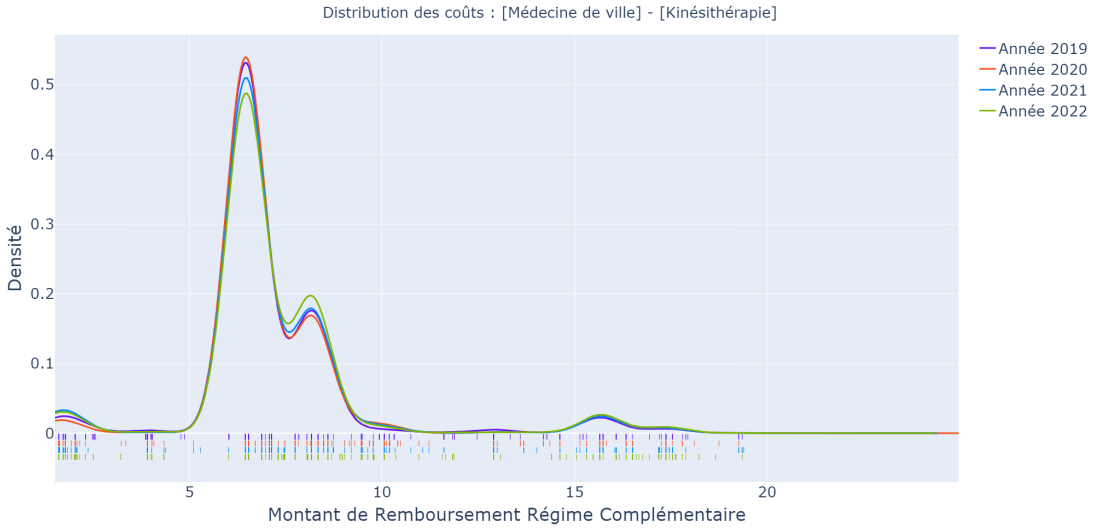


FIGURE 3.16 – Répartition des remboursements des lentilles de contact

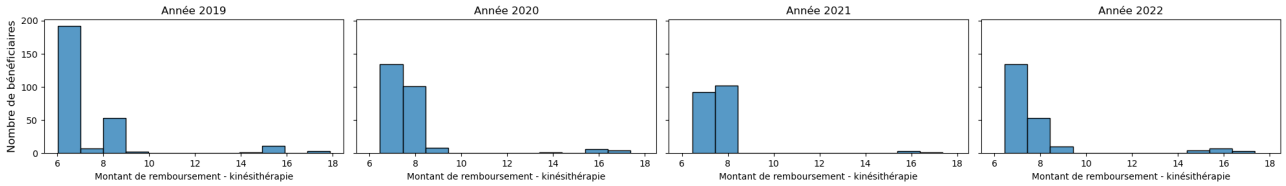
Le graphique 3.16 illustre la distribution des montants de remboursement complémentaire pour les séances de kinésithérapie sur plusieurs années, de 2019 à 2022. Les différentes courbes représentent les densités de probabilité des montants remboursés, avec des pics distincts qui reflètent la diversité des remboursements effectués.

Ces pics suggèrent l'existence de catégories distinctes dans les coûts de kinésithérapie, pouvant correspondre à différents types de traitements ou à des variations tarifaires. La concentration la plus dense de remboursements se situe entre 6 et 9 euros, ce qui est cohérent avec les montants

habituellement pris en charge après le remboursement de la Sécurité sociale.

La superposition des distributions pour les différentes années montre une stabilité dans la distribution des coûts, indiquant que les pratiques de remboursement et les coûts associés à la kinésithérapie ont été relativement constants sur la période observée. Cela suggère que les remboursements du régime complémentaire pour la kinésithérapie sont prévisibles, ce qui est une donnée essentielle pour l'estimation bayésienne des coûts dans le secteur de la santé.

FIGURE 3.17 – Montants de remboursement kinésithérapie - Entreprise Exemple



Le graphique 3.17 ci-dessus met en évidence la distribution des montants de remboursement pour la kinésithérapie au sein de l'entreprise analysée, répartie sur quatre années, de 2019 à 2022. On observe une répartition des montants qui reflète nos observations sur l'ensemble des entreprises. La majorité des remboursements se concentre entre 6 et 9 euros, ce qui correspond probablement au remboursement standard par séance de kinésithérapie. Il y a également un second groupe de remboursements, plus élevé, situé entre 14 et 18 euros, qui correspond principalement à deux séances ou possiblement trois séances effectuées le même jour pour le même bénéficiaire.

Pour modéliser cette distribution bimodale, nous envisagerons l'emploi d'un modèle de mélange, qui intègre deux distributions normales distinctes. La première représentera les remboursements les plus fréquents, avec une moyenne située entre 6 et 9 euros, tandis que la seconde modélisera le pic secondaire avec une moyenne autour de 16 euros. L'utilisation d'un tel modèle de mélange nous permettra de capturer la complexité de la distribution des coûts de kinésithérapie et d'affiner nos estimations des remboursements pour la kinésithérapie.

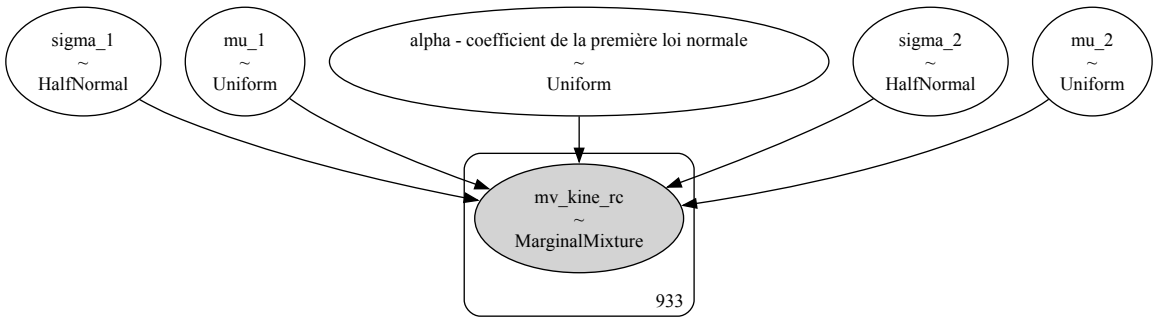


FIGURE 3.18 – Schéma du modèle des coût des séances de kinésithérapie

Le schéma 3.18 dépeint un modèle bayésien pour estimer les montants de remboursement de la kinésithérapie. Dans ce modèle, le paramètre α joue le rôle de coefficient de pondération pour la première distribution normale, dont l'espérance μ_1 est déterminée être entre 6 et 9 euros. Cette première distribution normale représente probablement le montant le plus couramment remboursé pour une

séance standard de kinésithérapie.

De plus, le complément à 1 du coefficient α, soit $1-\alpha$, sert de poids pour la deuxième distribution normale, dont l'espérance μ_2 est estimée autour de 16 euros. Cela pourrait correspondre à des situations où les coûts de kinésithérapie sont plus élevés, en raison de plusieurs séances réalisées le même jour.

Le modèle intègre également deux paramètres d'écart type σ_1 et σ_2 pour chacune des lois normales, qui sont modélisées comme des distributions demi-normales, indiquant que l'on ne considère que des valeurs positives pour ces variances. Cela est cohérent avec le fait que les écarts-types, mesurant la dispersion autour de la moyenne, ne peuvent être négatifs.

Les distributions a priori pour les espérances sont supposées uniformes, ce qui indique une absence de préférence pour une valeur spécifique avant l'observation des données. Cela permet au modèle de rester flexible et d'être mis à jour de manière informative avec les données observées.

En somme, ce schéma bayésien offre une structure pour combiner les informations sur les montants de remboursement fréquemment observés avec les cas moins courants mais plus coûteux, permettant ainsi une évaluation globale et nuancée des coûts de kinésithérapie pour l'entreprise étudiée.

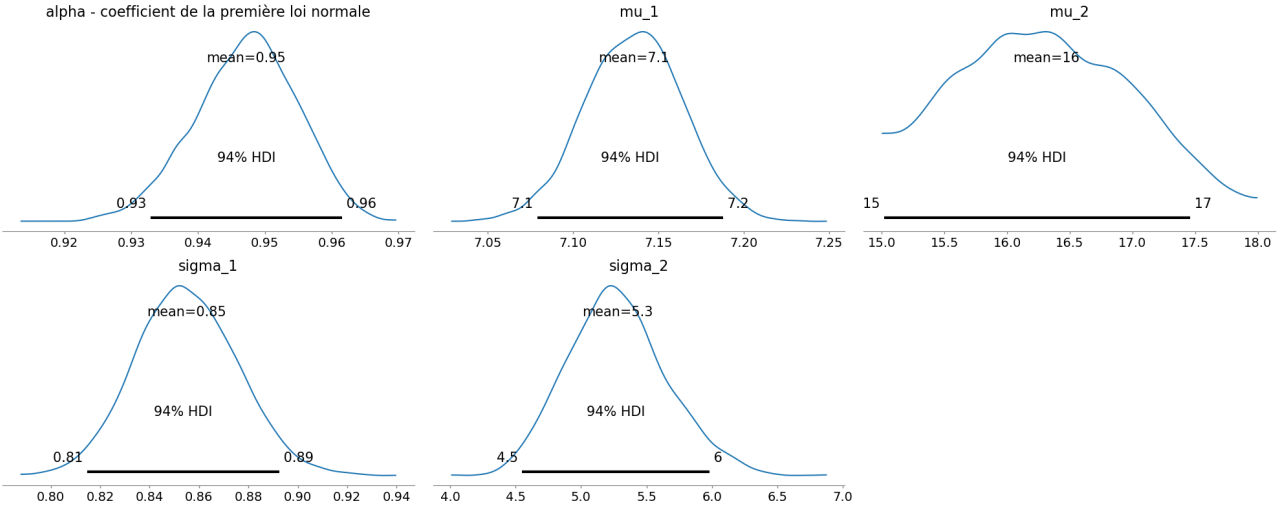


FIGURE 3.19 – Distribution a posteriori des paramètres

Les graphiques 3.19 représentent les distributions a posteriori pour les paramètres du modèle bayésien de mélange pour les remboursements de kinésithérapie. Le coefficient α, qui a une valeur autour de 0,95, indique que la première loi normale est privilégiée dans le modèle de mélange, avec 95% des observations s'alignant sur cette distribution. Cette loi a une espérance située entre 7,1 et 7,2 euros, et un écart type relativement faible d'environ 0,85, ce qui suggère une concentration élevée des remboursements autour de cette moyenne.

Pour la deuxième loi normale, qui représente un phénomène moins fréquent dans les données, l'espérance est estimée à environ 16 euros, mais avec un écart type plus large, reflétant une plus grande variabilité des montants remboursés. Cela peut être dû au faible nombre d'observations pour cette catégorie de remboursement, ce qui conduit à une plus grande incertitude autour de l'estimation de l'espérance et à un intervalle de crédibilité plus large.

L'interprétation de ces distributions a posteriori est cruciale pour comprendre la probabilité des différents montants de remboursement dans la population étudiée. Le modèle montre une forte certitude autour de la distribution principale des coûts de kinésithérapie et une plus grande incertitude pour les

cas moins fréquents et plus coûteux.

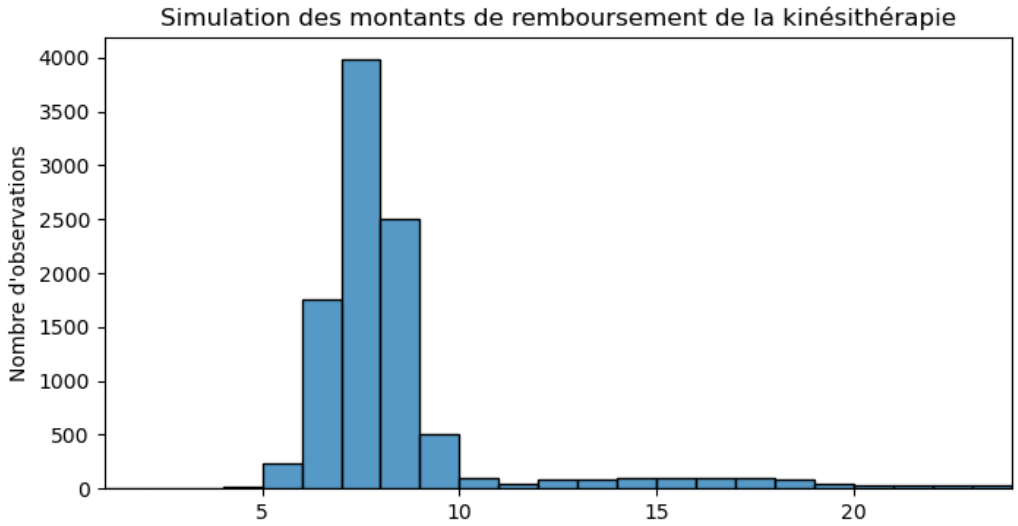


FIGURE 3.20 – Simulation de remboursement kinésithérapie - Entreprise Exemple

Le graphique 3.20 démontre les résultats d'une simulation des montants de remboursement pour la kinésithérapie au sein de l'entreprise analysée, en utilisant la distribution a posteriori des paramètres du modèle bayésien. Les résultats de la simulation semblent être en adéquation avec les observations précédentes tant pour l'ensemble des entreprises que pour l'entreprise étudiée.

La majorité des remboursements simulés se concentre autour de la tranche de 6 à 9 euros, reflétant la moyenne principale du modèle de mélange et corroborant l'hypothèse selon laquelle c'est le montant le plus couramment remboursé pour les séances de kinésithérapie. Il y a également quelques observations dans la tranche supérieure, vers 14 à 19 euros.

Ce modèle bayésien sert de fondement pour estimer les remboursements de kinésithérapie et est essentiel dans la détermination de la prime pure associée à ce type de prestation. En intégrant les résultats de nos simulations antérieures sur le nombre de séances, nous sommes en mesure de calculer une prime pure ajustée pour l'entreprise analysée. En adoptant une approche similaire à celle utilisée pour les lentilles de contact, nous calculons la moyenne des produits des montants simulés par le nombre de séances, ce qui nous permet d'obtenir une prime pure "a posteriori" pour l'entreprise concernée. Le résultat obtenu est présenté dans le tableau 3.3 ci-après.

TABLE 3.3 – Prime pure a posteriori par bénéficiaire - Kinésithérapie

Regroupement des actes	**Prime pure a posteriori**
Kinésithérapie	9,65 € par an par bénéficiaire

Regroupement d'actes - Chirurgie de l'oeil au laser

Dans ce troisième exemple, nous nous penchons sur un cas spécifique de regroupement d'actes : la chirurgie de l'œil au laser. C'est un exemple intéressant de simplicité, caractérisé par l'absence de sinistres observés pour l'entreprise étudiée. Comme dans le cas de garanties peu fréquentes telles que les prothèses auditives ou les véhicules handicapés, il n'est pas inhabituel de ne constater aucun sinistre passé en raison de la faible probabilité de leur occurrence.

Ce phénomène représente un défi significatif pour le secteur de l'assurance. L'emploi de méthodes statistiques classiques, telles que les modèles linéaires généralisés, peut entraîner une sous-estimation du risque en l'absence d'occurrences passées, même si le modèle est parfaitement calibré. À l'inverse, l'utilisation de moyennes basées sur l'ensemble de la population peut surévaluer le risque pour une entreprise individuelle.

L'approche bayésienne offre une perspective différente. Elle permet une "mise à jour" de la loi a priori en intégrant les données historiques, où l'absence de sinistre contribue substantiellement à cette mise à jour. De plus, le contexte de l'observation, qu'il s'agisse d'une petite entreprise sur une année, ou d'une plus grande entreprise sur quatre ans, influence fortement l'estimation du risque.

Nous allons donc examiner de plus près le cas spécifique de la chirurgie oculaire au laser, en appliquant la méthode bayésienne pour estimer ce risque au sein de l'entreprise étudiée qui n'a pas enregistré de cas antérieurs. Cette analyse nous permettra de comprendre comment gérer l'évaluation des risques lorsque les données historiques ne révèlent aucune réalisation concrète de ces risques.

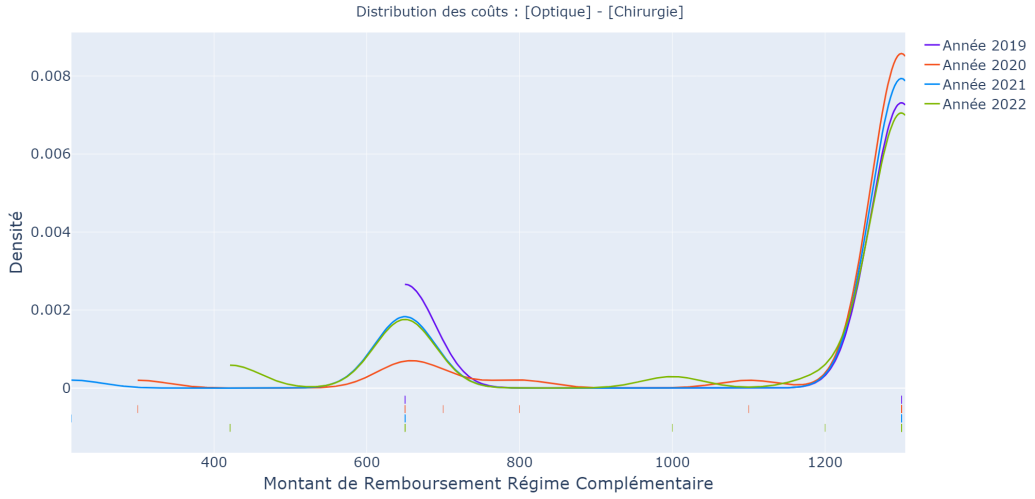


FIGURE 3.21 – Répartition des coûts de la chirurgie optique

Pour concentrer notre analyse, nous établirons une hypothèse concernant le montant des remboursements. Le graphique 3.21 illustre les montants remboursés pour les interventions de chirurgie optique à travers l'ensemble des entreprises du périmètre, l'examen de ces données suggère que les remboursements pour la chirurgie de l'œil au laser se stabilisent majoritairement autour de 1300 euros, correspondant au plafond de garantie fixé à 650 euros par œil. Bien que l'historique des données révèle des cas où les montants sont inférieurs à cette somme, dus soit à des interventions ne concernant qu'un seul œil, soit à des coûts opératoires moins élevés, ces occurrences sont relativement rares. Pour les besoins de notre étude, nous postulerons que le montant remboursé est uniformément de 1300 euros. Ainsi, notre analyse se focalisera davantage sur la fréquence de ces interventions, qui est un paramètre critique dans l'évaluation du risque pour une entreprise individuelle.

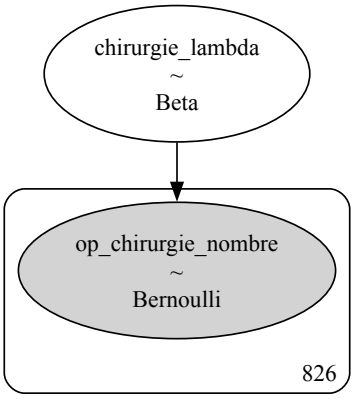


FIGURE 3.22 – Schéma du premier modèle des fréquences de la chirurgie optique

Le schéma 3.22 illustre un modèle statistique bayésien simple pour l'occurrence des opérations de chirurgie optique. Dans ce modèle, la survenue d'une chirurgie optique est décrite par une distribution de Bernoulli, ce qui signifie que chaque événement est indépendant et a deux issues possibles : il se produit ou ne se produit pas. Le paramètre lambda de cette distribution de Bernoulli, qui représente la probabilité de l'occurrence d'une opération de chirurgie optique, est lui-même modélisé par une distribution de la loi Bêta.

La distribution Bêta est particulièrement appropriée pour modéliser le paramètre d'une distribution de Bernoulli car elle est définie sur l'intervalle [0,1], ce qui correspond à l'ensemble des valeurs possibles pour une probabilité. De plus, la distribution Beta est flexible et peut prendre différentes formes en fonction de ses paramètres, ce qui lui permet de représenter une grande variété de croyances a priori sur la probabilité de l'événement étudié.

Le nombre 826 associé à la survenue d'une chirurgie optique indique le nombre total d'observations pris en compte dans le modèle. En résumé, ce schéma représente un modèle bayésien qui utilise des connaissances a priori, via la distribution Bêta, pour mettre à jour la probabilité de la chirurgie optique sur la base de données observées.

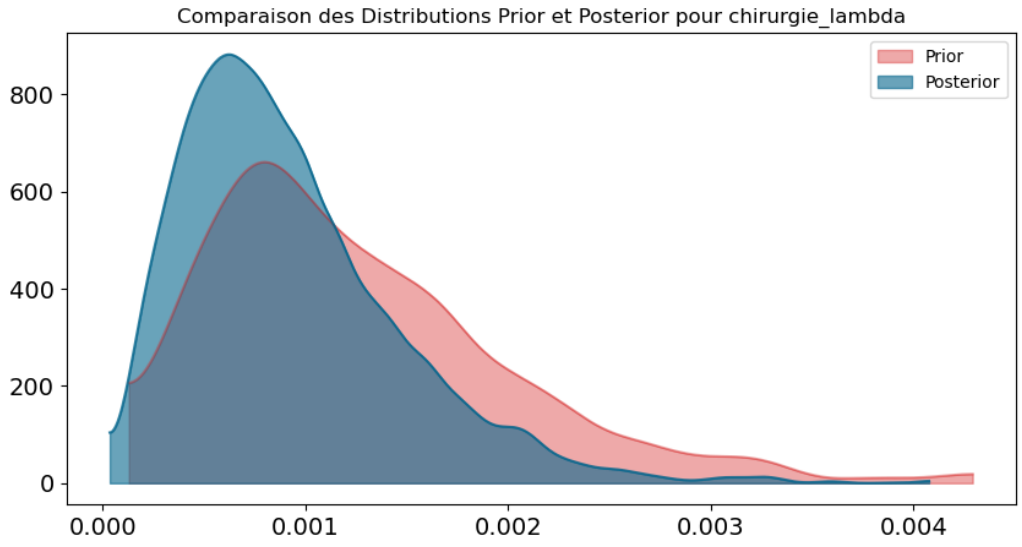


FIGURE 3.23 – Comparaison des distributions a priori et a posteriori

Le graphique 3.23 fournit une comparaison visuelle entre les distributions a priori et a posteriori pour le paramètre lambda dans le contexte de la chirurgie optique. La distribution a priori, représentée par la zone en rose, montre la croyance initiale quant à la probabilité de survenue d'une chirurgie optique avant d'observer les données. La distribution a posteriori, en bleu, ajuste cette croyance après avoir pris en compte les données observées, c'est-à-dire l'absence de chirurgies optiques réalisées au sein de l'entreprise étudiée.

La probabilité a posteriori de réalisation d'une chirurgie optique n'est pas nulle, comme le montre la distribution tracée par la courbe bleue, mais elle est nettement inférieure à la croyance a priori. Cela illustre le processus de mise à jour bayésienne : en l'absence d'événements observés (chirurgies optiques dans ce cas), la probabilité estimée de leur survenue diminue. Cependant, comme le nombre d'observations n'est pas négligeable, les données exercent une influence perceptible, ajustant la distribution a priori vers une distribution a posteriori qui reflète une probabilité plus faible de l'événement. Ce phénomène démontre l'efficacité de l'ajustement bayésien des estimations de probabilité en présence de nouvelles informations, même lorsque ces informations sont l'absence d'un événement.

Tout comme précédemment, la prime pure est déterminée à partir de la simulation de la probabilité d'occurrence de la chirurgie optique. Les résultats de ce calcul sont présentés dans le tableau 3.4 suivant.

TABLE 3.4 – Prime pure a posteriori par bénéficiaire - Chirurgie optique

Regroupement des actes	**Prime pure a posteriori**
Chirurgie optique	1,20 € par an par bénéficiaire

3.4 Synthèse et interprétation des résultats

En utilisant les méthodes exposées dans les sections précédentes, il est possible de déterminer la prime pure a posteriori pour une entreprise spécifique en additionnant les primes pures a posteriori correspondant aux différents regroupements d'actes. Nous avons sélectionné aléatoirement 30 entreprises disposant de données de 2019 à 2022. Nous avons ensuite exploité les informations des années 2019 à 2021 pour estimer les primes pures a posteriori.

Synthèse des résultats

Les cotisations hors taxes de l'année 2022 pour ces entreprises ont été extraites de la table des cotisations collectives. À l'aide de la formule suivante, nous sommes en mesure de calculer approximativement la prime pure par bénéficiaire pour chacune de ces 30 entreprises.

$$PP_{benef} = \frac{Cotis_{HT} * (1 - Taux_{frais})}{N_{benef}}$$

où N_{benef} est le nombre proportionnel de bénéficiaires de l'année 2022.

Pour évaluer l'efficacité de notre approche, nous avons effectué une comparaison entre les primes pures a posteriori, calculées à partir des données de 2019 à 2021, et les primes réglées déduisant les différents frais (frais d'assureur, frais de gestion, frais d'acquisition, etc.). Pour les contrats souscrits depuis plusieurs années, l'information relative à la prime pure n'est plus disponible. En raison des révisions tarifaires passées et de l'évolution des entreprises, une prime pure actuelle pour le même type d'entreprise n'est pas non plus pertinente. C'est pourquoi nous avons décidé de construire ici une approximation de la prime pure à partir des montants de cotisations hors taxes, en déduisant les différents frais.

Cette analyse nous a permis de constituer un tableau récapitulatif, présenté ci-dessous, qui met en lumière les résultats obtenus et offre une perspective sur la pertinence de notre méthode de calcul des primes pures a posteriori.

Table 3.5 – Tableau récapitulatif des résultats

Catégories de comparaison des primes pures	**Nombre d'entreprises**
Prime pure a posteriori inférieur à la prime approximative	23
Prime pure a posteriori supérieur, écart inférieur à 10%	5
Prime pure a posteriori supérieur, écart supérieur à 10%	2

Le tableau ci-dessus résume les résultats de la comparaison des primes pures entre les valeurs a posteriori calculées et les primes réglées déduisant les frais pour les entreprises étudiées. Il ressort que la majorité des entreprises ont une prime pure a posteriori inférieure à la prime pure approximative par bénéficiaire, ce qui est conforme aux attentes, car le périmètre de l'étude concerne un portefeuille rentable depuis des années. De plus, la prime pure approximative inclut non seulement le coût des sinistres, mais aussi une marge technique destinée à couvrir l'incertitude liée à l'évaluation des engagements futurs.

Seule une minorité des entreprises ont une prime pure a posteriori qui dépasse la prime réglé déduisant les frais, avec des écarts inférieurs ou supérieurs à 10%. Ces écarts peuvent découler de plusieurs facteurs, y compris les spécificités de l'entreprise, les tendances du marché, ou la structure des coûts et des marges. Une analyse détaillée de ces différences est abordée dans la section suivante pour fournir une compréhension plus fine de ces dynamiques.

Interprétations des résultats

Avec la synthèse de la comparaison des primes pures, nous constatons que les résultats obtenus correspondent à nos attentes. Cependant, afin de mieux comprendre la performance et le comportement des modèles bayésiens, nous avons examiné en détail les estimations des entreprises, leurs données historiques et leurs primes pures actuelles, ce qui nous a permis de tirer plusieurs conclusions pouvant être résumées comme suit.

- **Impact des fluctuations récentes sur l'estimation des primes pures** :

Notre méthode de calcul de la prime pure a posteriori s'est largement appuyée sur les données accumulées entre 2019 et 2021, une période marquée par une volatilité considérable dans le secteur de l'assurance santé.

Cette volatilité a été principalement provoquée par la pandémie de Covid-19, qui a engendré une diminution notable de la fréquence des soins de santé en 2020. Même si l'effet global sur l'ensemble des entreprises a été relativement maîtrisé et quantifiable, l'impact sur des entreprises individuelles a varié, présentant des conséquences plus ou moins prononcées. Un exemple notable est celui des entreprises dont les salariés ont réduit leur recours aux soins en 2020, entraînant une baisse de la prime pure a posteriori calculée, bien que cet effet puisse ne pas être durable. Même si un effet de rattrapage mineur a été observé en 2021, cela a accru l'incertitude de nos estimations.

De surcroît, cette période a coïncidé avec le lancement de la réforme 100% santé, visant à améliorer l'accès aux soins optiques, dentaires et auditifs. Alors que l'impact sur l'ensemble du portefeuille d'assurance est resté stable, nos analyses révèlent des effets significatifs à l'échelle individuelle, surtout pour les petites et moyennes entreprises. Nous avons observé des changements de comportement notables en matière de soins dentaires en 2020 et 2021 par rapport à 2019, ce qui a conduit à une augmentation de la prime pure a posteriori pour certaines entreprises. Ces observations soulignent la complexité de l'estimation tarifaire dans un contexte fluctuant et mettent en évidence la nécessité d'une approche nuancée pour appréhender l'impact de ces changements sur les primes pures d'assurance santé.

- **Le rôle clé de la sélection des distributions a priori** :

L'adoption de distributions a priori adéquates et la modélisation précise des distributions a posteriori sont cruciales pour l'exactitude de nos estimations. Comme illustré par nos exemples, un choix judicieux de ces distributions ne se limite pas à fournir des estimations précises, il est aussi essentiel pour aligner ces estimations sur nos prévisions et attentes. Cependant, il convient de noter que les fréquences et les montants de remboursement sont sujets à une large variabilité, rendant le modèle sensible à cette dispersion.

D'autre part, l'utilisation d'une distribution a priori non informative peut sembler simplifier le modèle et accélérer le processus de calcul. Néanmoins, cette approche présente le risque d'être trop influencée par les données spécifiques à une entreprise, sans forcément parvenir à un équilibre entre les tendances générales observées au sein de l'ensemble des entreprises et les particularités d'une entreprise donnée. La complexité accrue du modèle peut certes augmenter le temps de calcul et sa sensibilité aux valeurs extrêmes ou aberrantes, mais elle peut être nécessaire pour capturer la réalité sous-jacente de manière plus fidèle.

Dans le cadre de notre étude, nous avons sélectionné des distributions a priori qui se conjuguent au mieux avec notre compréhension des diverses catégories d'actes de soins. Cette sélection s'est principalement orientée vers des distributions a priori informatives. Toutefois, il est important de noter que ce choix peut parfois limiter la précision des prédictions. En effet, une distribution a priori particulièrement forte peut atténuer l'influence des données historiques, un effet particulièrement marqué pour les petites entreprises qui disposent de moins de données.

D'un point de vue prédictif, une démarche pragmatique consiste à tester plusieurs distributions a priori et à évaluer leur efficacité respective. Cette approche permet de déterminer la distribution

a priori la plus adéquate, qui reflète fidèlement nos connaissances et qui est susceptible de s'adapter aisément aux différentes tailles d'entreprises. Dans le prochain chapitre, nous explorerons certaines méthodes et métriques pour évaluer et comparer ces distributions a priori, afin d'optimiser nos choix et de renforcer la pertinence de nos modèles prédictifs.

- **L'impact de la granularité dans la modélisation bayésienne** :

La finesse avec laquelle nous appliquons notre modèle a également démontré son importance. Pour cette étude, nous avons opté pour la modélisation spécifique à une quarantaine de catégories de soins par entreprise. Cette approche détaillée par catégorie et par entreprise affine notre compréhension des mécanismes des modèles bayésiens. Néanmoins, un regroupement plus global des actes pourrait simplifier la création des modèles tout en augmentant leur robustesse. Toutefois, une telle agrégation demanderait des analyses encore plus poussées pour rester pertinentes.

Appliquer ces modèles aux entreprises nous fournit des aperçus détaillés sur chacune d'elles. Cependant, notre étude étant limitée principalement à des petites et moyennes entreprises, le manque de significativité dû à des nombres d'observations limités ne nous permet pas toujours de différencier clairement une entreprise d'une autre. En conséquence, les distributions a posteriori tendent à rester proches des distributions a priori. Ces méthodes pourraient s'avérer plus adaptées et fournir des résultats plus distinctifs pour des entreprises de plus grande taille.

- **Considération des effets temporels dans l'estimation bayésienne** :

L'influence des variations annuelles est également un facteur significatif, comme nous l'avons souligné lors de l'estimation bayésienne de la fréquence annuelle des lentilles de contact où les paramètres peuvent fluctuer d'année en année, révélant des tendances spécifiques. Dans le cadre de cette recherche, nous n'avons pas intégré les effets temporels dans nos modèles des paramètres, mais il est tout à fait possible de les incorporer dans une approche bayésienne en utilisant des méthodes plus avancées telles que les processus gaussiens, sujet que nous aborderons dans le prochain chapitre.

Conclusion

Dans ce chapitre, nous avons d'abord exploré les principes fondamentaux de la tarification en assurance santé collective, en mettant l'accent sur la distinction entre la tarification a priori et a posteriori. Cette exploration a permis de poser les bases théoriques nécessaires à la compréhension des mécanismes sous-jacents de la tarification dans le domaine de l'assurance santé, ainsi que des concepts clés des deux approches de tarification, a priori et a posteriori. Elle a également mis en lumière les contextes et les avantages de chacune de ces approches, tout en soulignant l'importance d'adapter les méthodes de tarification aux spécificités de chaque enjeu.

Ensuite, le chapitre s'est orienté vers la présentation du modèle bayésien, en introduisant des notions essentielles telles que la distribution a priori, la vraisemblance et la distribution a posteriori, tout en intégrant des techniques avancées comme les méthodes de Monte-Carlo par chaînes de Markov, le modèle de mélange et le modèle hiérarchique bayésien. Cette section a été cruciale pour démontrer la manière dont les méthodes bayésiennes peuvent être appliquées pour intégrer les données historiques

aux croyances initiales, afin d'affiner les estimations de tarification. L'adoption de ces approches statistiques a permis de mieux saisir la complexité et la variabilité inhérentes aux données du secteur de l'assurance santé.

Enfin, nous avons étudié l'application pratique des modèles bayésiens à travers trois exemples de regroupements d'actes, illustrant concrètement comment ces modèles peuvent être appliqués dans des situations réelles. Ces trois exemples représentent trois types de soins de santé : la fréquence des soins peut être occasionnelle, récurrente ou rare, et le montant de remboursement peut être moyen, faible ou élevé. De la même manière, nous avons calculé les primes pures pour tous les regroupements d'actes de soins, afin de construire la prime pure pour une entreprise. L'étude s'est ensuite concentrée sur un échantillon de 30 entreprises pour appliquer la méthode de tarification bayésienne. L'analyse des résultats a fourni des indications sur l'efficacité de l'approche bayésienne, en particulier dans le contexte de la tarification a posteriori pour les entreprises de taille importante du secteur de l'assurance santé collective.

Chapitre 4

Approfondissement et perspective

Dans ce chapitre, nous plongerons dans l'exploration de sujets avancés, avec un accent particulier sur la tarification a posteriori, tout en adoptant et en approfondissant l'approche bayésienne. Cette section vise à élargir les horizons de la tarification actuarielle, en intégrant des concepts et des méthodologies avancées pour enrichir notre compréhension et notre application de l'actuariat dans un monde de plus en plus complexe et interconnecté.

Bien que ces sujets n'aient pas été explorés en détail dans les chapitres précédents, principalement en raison de leur complexité, leur importance et leur utilité dans le domaine de l'actuariat moderne ne peuvent être sous-estimées. Ces thématiques représentent non seulement l'avant-garde de la recherche actuelle en tarification actuarielle et statistique bayésienne, mais offrent également des perspectives pour l'avenir de la profession.

4.1 Modèles hiérarchiques bayésiens

Dans le chapitre précédent, nous avons introduit les modèles hiérarchiques bayésiens de manière concise, mettant en évidence leur potentiel considérable dans l'application des modèles bayésiens. Ces modèles se distinguent par leur capacité à structurer les données de manière hiérarchique, permettant une analyse plus nuancée et détaillée.

Concrètement, un modèle hiérarchique bayésien organise les paramètres et les données à différents niveaux, intégrant les variations à chaque échelon. Cette approche multi-niveaux permet de capturer des relations complexes au sein des données, en traitant à la fois les variations globales et les spécificités locales. En incorporant les principes de l'inférence bayésienne, ces modèles offrent une robustesse remarquable pour la mise à jour des probabilités et des estimations de paramètres en fonction de nouvelles données, rendant ainsi leur application en tarification a posteriori particulièrement pertinente et efficace.

Au cours de cette étude, nous avons principalement exploité un volume substantiel de données détaillées sur les prestations, englobant plus de 10 millions de lignes de données. Notre analyse s'est concentrée sur des aspects essentiels tels que la population couverte, les types de soins, et les montants de remboursement correspondants. Parallèlement, nous avons également eu accès aux données démographiques sur les entreprises, y compris les âges moyens et les répartitions par genre. De plus, nous avons disposé d'informations complémentaires concernant ces entreprises, telles que leur secteur d'activité et leur localisation géographique, cette dernière étant définie par le code de département.

Traditionnellement, ce type de données est souvent mobilisé pour établir la prime pure a priori, c'est-à-dire avant la souscription du contrat et avant d'avoir une connaissance précise du client, afin de déterminer le niveau de risque associé à différents groupes d'entreprises. Cependant, ces informations jouent également un rôle crucial dans le cadre de la tarification a posteriori lorsqu'on adopte une approche bayésienne. En effet, grâce à l'utilisation des modèles hiérarchiques bayésiens, nous pouvons affiner et calibrer les paramètres de nos modèles avec une plus grande robustesse. Ces modèles facilitent l'intégration de données à multiples niveaux, allant des détails spécifiques aux entreprises, tels que leurs caractéristiques démographiques, sectorielles ou géographiques, jusqu'aux tendances globales

perceptibles au sein de divers groupes et secteurs, ce qui permet d'affiner l'estimation des risques de manière plus précise et détaillée.

Dans la continuité de notre exploration des modèles bayésiens, nous revisitons le modèle que nous avons développé précédemment pour estimer la fréquence d'achat de lentilles de contact. En intégrant les concepts des modèles hiérarchiques bayésiens, nous avons la possibilité d'enrichir notre modèle initial avec des couches supplémentaires de détails. Cela inclut l'intégration de données supplémentaires telles que la tranche d'âge moyen et la répartition par genre des entreprises, ainsi que le code de département, offrant ainsi une modélisation plus complète et nuancée.

L'adoption de cette approche hiérarchique permet de reconnaître et d'analyser les variations dans les comportements d'achat non seulement au niveau individuel, mais aussi en fonction de facteurs démographiques et géographiques plus larges des entreprises. Par exemple, en analysant les données à travers les âges moyens et les répartitions par genre, nous pouvons identifier des tendances spécifiques et des préférences qui varient selon ces catégories. De même, l'inclusion des codes de département nous aide à comprendre comment la localisation géographique des entreprises influence les habitudes d'achat.

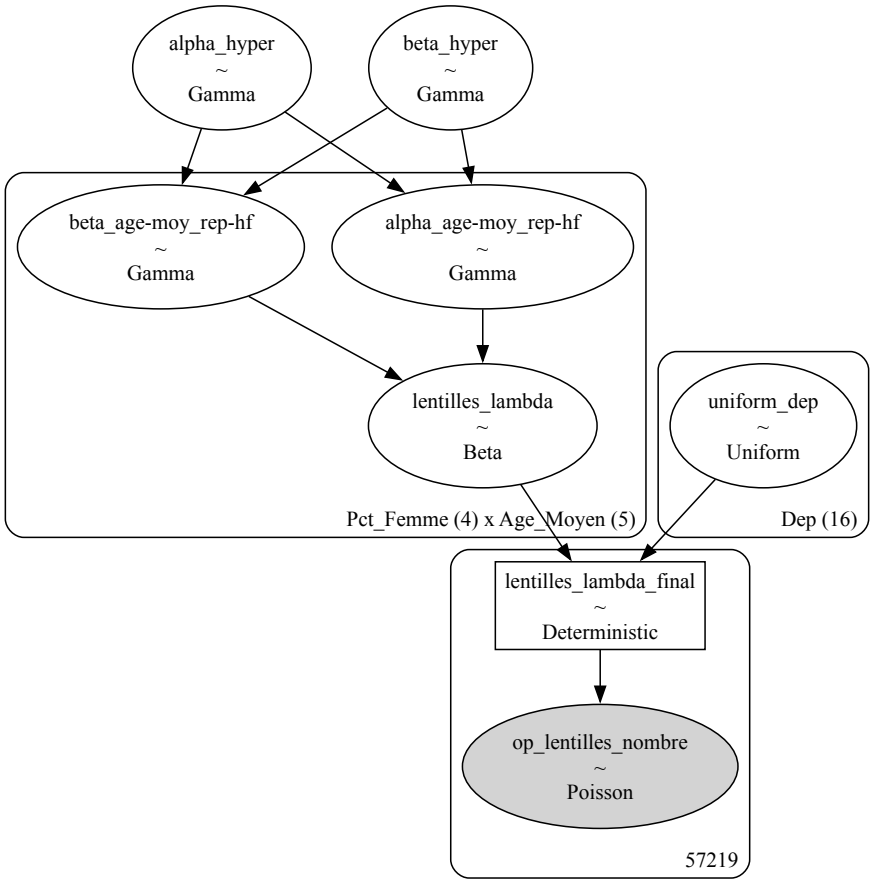


FIGURE 4.1 – Schéma du modèle hiérarchique bayésien - Lentilles de contact

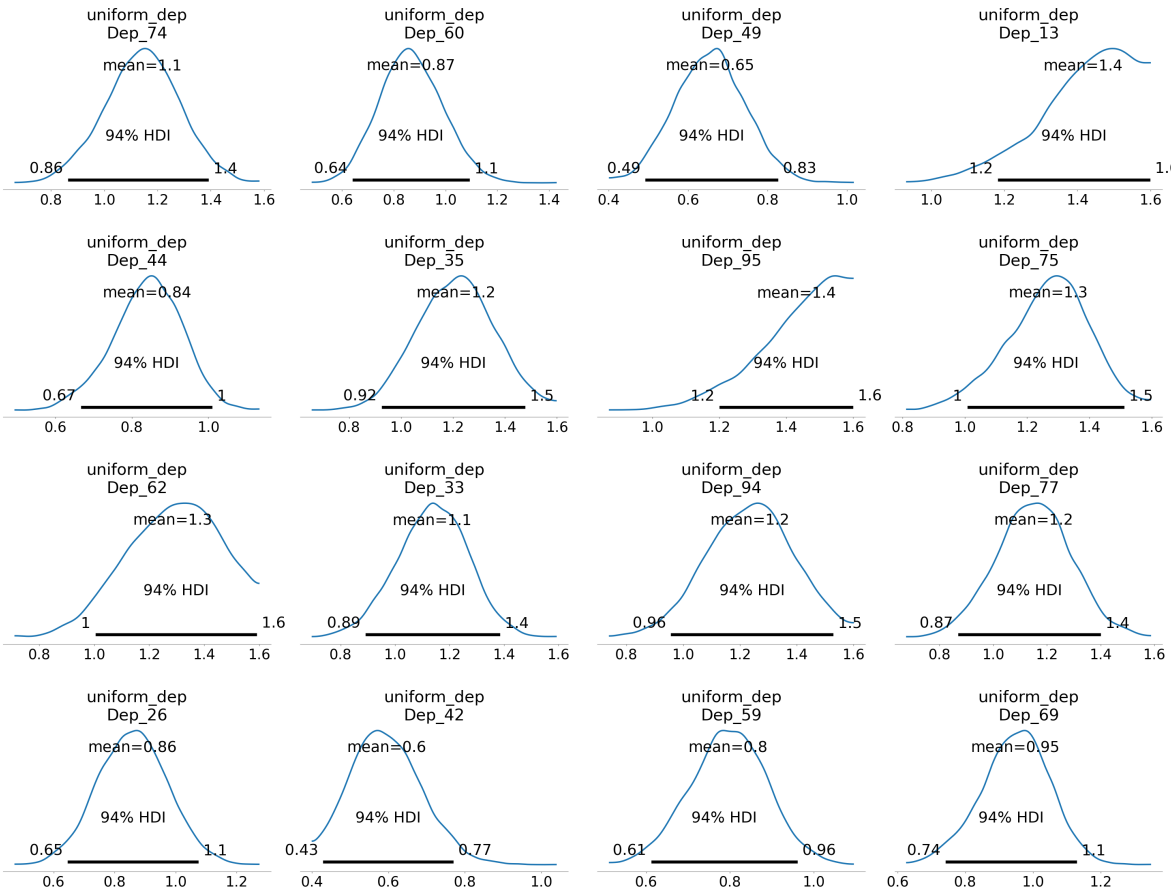
Le schéma 4.1 illustre un modèle hiérarchique bayésien destiné à estimer la fréquence d'achat de lentilles de contact, intégrant des variables démographiques et géographiques. En tête du schéma, nous trouvons deux hyperparamètres, *alpha_hyper* et *beta_hyper*, chacun suivant une distribution Gamma proprement définie. Ces hyperparamètres capturent les informations globales sur l'ensemble de la population, servant de fondement à la hiérarchie du modèle.

En descendant dans la hiérarchie, pour chaque combinaison spécifique d'âge moyen et de répartition par genre, nous disposons de paramètres *alpha_ age-moy_ rep-hf* et *beta_ age-moy_ rep-hf* distincts. Ces paramètres sont spécifiques à chaque sous-groupe démographique. Dans le cadre de cette étude, et afin de simplifier l'interprétation des résultats, nous avons créé 4 groupes de répartition par genre, représentés par des tranches de pourcentage de femmes, ainsi que 5 tranches d'âge moyen des entreprises. En pratique, l'utilisation d'un plus grand nombre de tranches rendrait le modèle plus précis. Le paramètre *lentilles_ lambda*, est ensuite dérivé de ces deux paramètres *alpha_ age-moy_ rep-hf* et *beta_ age-moy_ rep-hf* pour chaque combinaison de tranche d'âge moyen et de type de répartition par genre, suivant une distribution Beta, reflétant ainsi la variabilité spécifique à chaque sous-groupe.

En parallèle, l'information géographique liée au département [1] est incorporée à l'aide du paramètre *uniform_ dep* d'une distribution a priori uniforme, qui permet d'ajuster le paramètre lambda en fonction de la localisation. Cette loi uniforme, qui varie entre 0,4 et 1,6, sert à ajuster le paramètre lambda initial en le multipliant pour obtenir le paramètre lambda final. Ce dernier paramètre est déterministe, ce qui signifie qu'il est calculé directement à partir des valeurs précédentes sans variation supplémentaire.

Finalement, le modèle aboutit à la variable *op_ lentilles_ nombre*, qui suit une distribution de Poisson. Cette variable est utilisée pour estimer le nombre annuel attendu d'achats de lentilles de contact. Ce modèle hiérarchique bayésien permet ainsi une estimation plus nuancée et robuste de la fréquence des soins, en tenant compte des effets démographiques et géographiques.

FIGURE 4.2 – Distribution des paramètres - Département



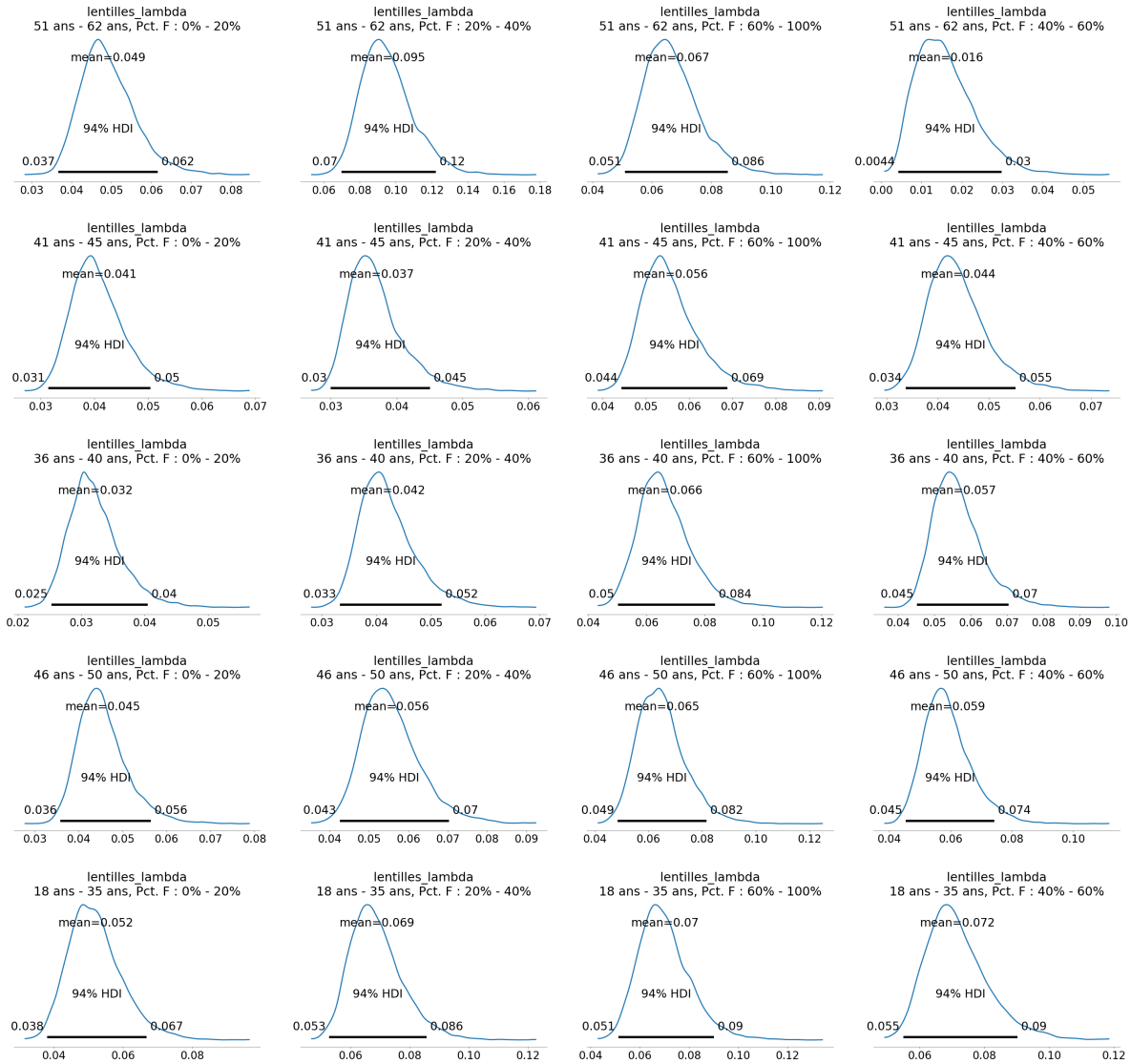
Le graphique 4.2 illustre les distributions des paramètres *uniform_ dep* pour différents départe-

1. Pour simplifier l'affichage, seuls 16 départements sont sélectionnés dans cet exemple.

ments. On observe que les départements 95 et 13 se démarquent avec des valeurs moyennes plus élevées, suggérant une augmentation de la fréquence estimée des achats dans ces zones. En contraste, les départements 42 et 49 présentent des valeurs moyennes inférieures, suggérant une réduction potentielle du coefficient lambda, ce qui pourrait indiquer un taux d'achat réduit pour la population résidant dans ces départements.

Ces distributions offrent une vision précise de l'influence géographique sur le comportement d'achat et sont cruciales pour l'élaboration d'un modèle de tarification à la fois robuste et précis. En plus de renforcer notre modèle, ces données nous permettent de vérifier la validité de nos hypothèses de tarification a priori, ce qui est indispensable pour confirmer nos attentes en termes de rentabilité et d'ajuster notre stratégie en conséquence.

FIGURE 4.3 – Distribution des paramètres - Démographie



Le graphique 4.3 illustre les distributions des paramètres *lentilles_ lambda* pour différents types d'entreprises, en utilisant une loi de Poisson où ce paramètre représente l'espérance, indiquant la fréquence prévue d'achat de lentilles de contact. On observe des disparités marquées entre les différents groupes de répartition par genre ainsi qu'au sein des différentes tranches d'âge moyen. Pour la plupart

des groupes de répartition par genre, la fréquence d'achat prévue est la plus élevée dans la tranche d'âge moyen de 18 à 35 ans. À tranche d'âge équivalente, on observe que la répartition par genre influence de manière significative la distribution a posteriori du paramètre *lentilles_lambda*. L'intégration des données démographiques des entreprises rend ainsi le modèle bayésien plus robuste, en permettant une meilleure adaptation aux caractéristiques spécifiques de chaque groupe.

Cette différenciation démographique, intégrée dans notre modèle bayésien, augmente la robustesse de notre modèle. Elle offre aussi une clarté d'interprétation qui nous aide à mieux saisir et à apprécier les variations comportementales entre les groupes. Ces aperçus fournissent des informations d'une valeur précieuse, non seulement pour la tarification a posteriori, où l'adaptation des prix est basée sur des comportements observés, mais aussi pour la tarification a priori, qui s'appuie sur des estimations de risque avant l'émergence de données concrètes de sinistralité.

En conclusion de notre exploration des modèles hiérarchiques bayésiens, il est clair que leur intégration dans la tarification a posteriori ouvre de nouvelles perspectives pour une compréhension plus fine et une prédiction plus précise dans le domaine de l'assurance. Ces modèles, par leur capacité à intégrer divers niveaux d'information, permettent une segmentation et une analyse détaillée des risques couverts. Ainsi, les assureurs peuvent ajuster les tarifs avec une granularité qui reflète fidèlement le risque inhérent à chaque groupe spécifique.

L'application de modèles hiérarchiques bayésiens à la tarification a posteriori présente plusieurs avantages. Le premier est la personnalisation : les tarifs ne sont plus uniformément appliqués à un grand ensemble d'entreprises similaires, mais sont ajustés selon des critères précis, rendant la tarification plus équitable et compétitive. Ensuite, il y a l'adaptabilité : ces modèles sont conçus pour s'actualiser en intégrant de nouvelles données, ce qui les rend idéaux dans un monde où les tendances de consommation et les risques évoluent rapidement. Enfin, ils favorisent la transparence et la compréhension des facteurs de risque, un atout pour les compagnies d'assurance qui cherchent à communiquer plus efficacement avec leurs clients et régulateurs.

Cependant, les modèles hiérarchiques bayésiens sont plus complexes, leur conception et leur interprétation nécessitent davantage d'analyses statistiques ainsi que de tests supplémentaires pour garantir une modélisation adéquate. De plus, les calculs impliqués sont souvent intensifs et peuvent nécessiter une puissance de traitement élevée, surtout lorsqu'ils sont appliqués à de grands ensembles de données. Ces contraintes computationnelles peuvent limiter l'agilité dans l'ajustement des modèles et dans la réponse aux nouvelles informations.

Malgré ces défis, l'adoption des modèles hiérarchiques bayésiens en tarification a posteriori ouvre de nouvelles possibilités dans le domaine de l'actuariat. Avec une approche minutieuse et une attention particulière à la robustesse et à l'interprétation, ils représentent un outil puissant pour améliorer les stratégies de tarification dans l'industrie de l'assurance, en alliant précision analytique et agilité décisionnelle.

4.2 Applications des processus Gaussiens dans les modèles bayésiens

Dans le chapitre précédent, nous avons exploré l'exemple de l'estimation de la fréquence d'achat de lentilles de contact pour une entreprise donnée, couvrant la période de 2019 à 2022. Au cours de ces quatre années, nous avons observé une tendance à la baisse du paramètre lambda. Toutefois, les méthodes bayésiennes abordées jusqu'à présent s'avèrent limitées pour modéliser de telles variations

annuelles. Pour remédier à cela, il est judicieux de se tourner vers des techniques plus avancées, telles que l'application des processus gaussiens dans les modèles bayésiens.

Les processus gaussiens offrent un cadre puissant et flexible pour modéliser des données complexes et évolutives. Ils sont particulièrement pertinents pour capturer des tendances et des variations temporelles, comme celles observées dans notre étude. Un processus gaussien est un modèle probabiliste qui étend la notion de distributions gaussiennes à des fonctions infinies. En d'autres termes, il permet de définir une distribution sur un espace de fonctions, rendant possible la modélisation de données séquentielles ou temporelles.

- **Définition de base** :

Un processus gaussien est une collection de variables aléatoires, dont toute sous-collection finie suit une distribution multivariée gaussienne. Mathématiquement, un processus gaussien sur un espace X est défini par une fonction moyenne $m(x)$ et une fonction de covariance (aussi appelée noyau) $k(x, x')$, où $x, x' \in X$.

- **Formulation mathématique** :

Soit $f(x)$ une fonction de l'espace X vers les réels. On dit que $f(x)$ est distribuée selon un processus gaussien si pour tout ensemble $\{x_1, x_2, ..., x_n\} \subset X$, le vecteur $[f(x_1), f(x_2), ..., f(x_n)]^T$ suit une distribution gaussienne multivariée. Formellement, cela s'écrit comme :

$$f\left(x\right) \sim \mathcal{GP}(m(x), k(x, x'))$$

où $m(x)$ est la fonction moyenne, $k(x, x')$ est la fonction de covariance.

- **Fonction de covariance** :

La fonction de covariance, ou noyau, est cruciale dans la définition d'un processus gaussien. Elle détermine les propriétés de régularité et de continuité de la fonction aléatoire $f(x)$. Des exemples courants de fonctions de covariance incluent le noyau exponentiel quadratique et le noyau périodique.

- **Application dans les modèles bayésiens** :

Dans un cadre bayésien, les processus gaussiens sont utilisés comme des distributions a priori sur des fonctions. Lors de l'observation de données, ces distributions a priori sont mises à jour pour former des distributions a posteriori, qui capturent l'incertitude restante sur la fonction inconnue $f(x)$. Cela permet d'effectuer des prédictions pour de nouvelles entrées x_* , en tenant compte à la fois de l'incertitude et de la corrélation présumée avec les données observées.

- **Prédiction avec processus gaussien** :

Supposons que l'on dispose d'un ensemble d'observations Y aux points X. Pour prédire la valeur en un nouveau point x_*, on utilise la distribution conditionnelle des processus gaussiens, qui est elle-même gaussienne :

$$f(x_*)|X, Y, x_* \sim \mathcal{N}(\bar{f}(x_*), var(f(x_*)))$$

où $\bar{f}(x_*)$ et $var(f(x_*))$ sont déterminés par les données observées et la fonction de covariance.

L'une des forces des processus gaussiens réside dans leur capacité à intégrer des informations a priori sur la nature des données. Par exemple, si l'on s'attend à une certaine régularité, à des motifs répétitifs, ou encore à une tendance temporelle spécifique dans les données, les processus gaussiens peuvent être ajustés pour refléter ces attentes. Cette flexibilité les rend idéaux pour une variété d'applications, allant de la prédiction de séries temporelles à l'optimisation de fonctions complexes.

Dans le contexte des modèles bayésiens, l'application des processus gaussiens permet de surmonter certaines limitations des approches conventionnelles. Par exemple, ils peuvent capturer avec précision des tendances non linéaires et des interactions complexes entre variables, qui pourraient autrement passer inaperçues. En outre, les processus gaussiens offrent une estimation naturelle de l'incertitude, un élément clé dans la prise de décision basée sur les données.

En résumé, l'incorporation des processus gaussiens dans le cadre des modèles bayésiens représente une avancée significative pour l'analyse et la prédiction de phénomènes complexes. Cette méthode se révèle particulièrement efficace dans des cas tels que l'étude de la variation de la fréquence d'achat de lentilles de contact, où les tendances et les dynamiques sous-jacentes sont subtiles et multidimensionnelles. L'emploi de processus gaussiens enrichit notre compréhension en offrant une analyse plus nuancée et détaillée, ce qui conduit à des prédictions plus précises et informatives. En outre, cette approche facilite l'identification de modèles sous-jacents et de corrélations, permettant ainsi une interprétation plus approfondie des données et une meilleure prise de décision basée sur des analyses prédictives de haute qualité.

4.3 Évaluation et validation des distributions a priori

Dans le cadre de notre exploration des modèles bayésiens, nous avons identifié l'importance du choix de la distribution a priori, tant d'un point de vue théorique que pratique. Cette section est consacrée à la présentation de divers critères clés destinés à comparer et évaluer ces distributions a priori. Nous aborderons des méthodes et des outils d'évaluation, permettant une appréciation approfondie de l'impact et de la pertinence des distributions a priori choisies dans le contexte de nos modèles bayésiens.

- **La divergence de Kullback-Leibler** :

La divergence de Kullback-Leibler *(KL)* est un outil statistique essentiel pour mesurer la différence entre deux distributions probabilistes, souvent utilisée dans le cadre des modèles bayésiens. La divergence de Kullback-Leibler entre deux distributions de probabilité P et Q est définie comme suit :

Pour des distributions discrètes :

$$D_{KL}(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)}$$

Pour des distributions continues :

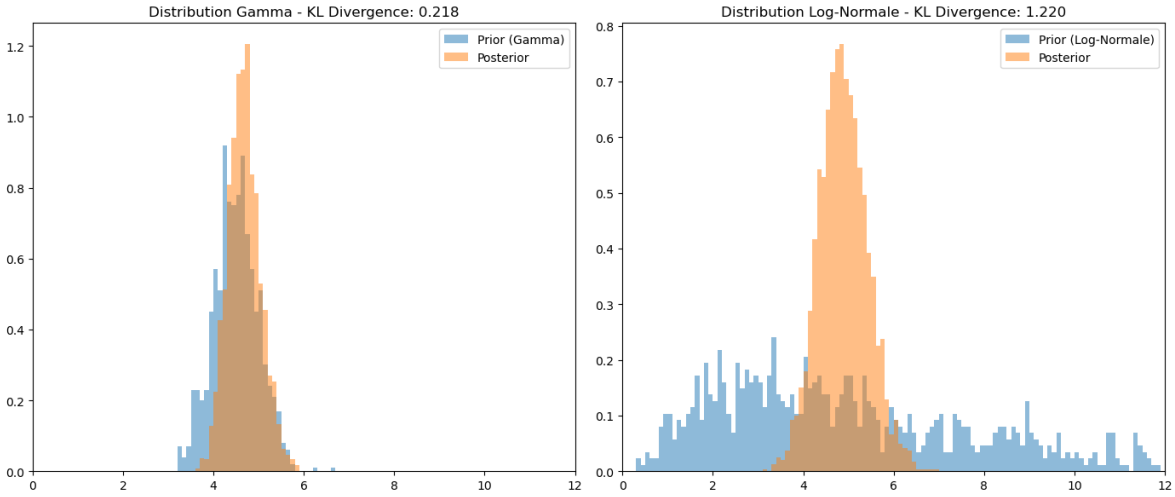
$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) log \frac{p(x)}{q(x)} dx$$

où $P(x)$ et $Q(x)$ (ou $p(x)$ et $q(x)$ pour les distributions continues) représentent les fonctions de masse de probabilité ou de densité de probabilité de P et Q, respectivement.

La divergence de Kullback-Leibler mesure l'information perdue lorsque Q est utilisée pour approximer P. Une valeur de 0 indique que les deux distributions sont identiques. Pour évaluer si la distribution a posteriori obtenue après l'analyse bayésienne s'écarte significativement d'une distribution de référence (par exemple, une distribution a priori non informative ou une distribution obtenue dans des études précédentes), la divergence de Kullback-Leibler peut être calculée entre ces deux distributions.

Dans les cas où différentes distributions a priori sont utilisées, la divergence de KL permet de mesurer l'impact de ces choix sur la distribution a posteriori. En calculant la divergence de KL entre les distributions a posteriori résultant de différentes lois a priori, on peut évaluer l'influence de l'hypothèse a priori sur l'inférence bayésienne.

FIGURE 4.4 – Comparaison des distributions a priori : informative et non-informative



Dans le graphique 4.4, nous comparons deux distributions a priori pour estimer le paramètre d'une loi de Poisson. La première est une distribution a priori Gamma informative, qui intègre des connaissances globales préexistantes, et la seconde est une distribution a priori log-normale non-informative, ne contenant pas d'information préalable spécifique.

Nous constatons que la divergence de Kullback-Leibler pour la distribution a priori Gamma est de 0,218, ce qui indique une faible différence entre la distribution a priori et la distribution a posteriori. Cela est conforme aux attentes, car l'a priori contient déjà des informations pertinentes, rendant les ajustements postérieurs relativement minimes. En revanche, la divergence KL pour la distribution log-normale non-informative est plus élevée, avec une valeur de 1,220. Cela signifie que la distribution a priori log-normale diverge plus fortement de la distribution a posteriori, reflétant le manque d'information initial dans l'a priori.

Enfin, nous observons que la distribution a posteriori issue de l'a priori Gamma présente une variance plus faible que celle obtenue à partir de la distribution log-normale. Cela montre que l'a priori informative renforce notre confiance dans la distribution a posteriori, un effet particulièrement pertinent dans des contextes tels que l'estimation des sinistres, où des connaissances solides préexistantes peuvent être exploitées pour améliorer la précision de l'estimation.

En conclusion, la divergence de Kullback-Leibler fournit un moyen quantitatif de comparer et d'évaluer des distributions dans le cadre de l'inférence bayésienne, en mesurant l'écart entre les distributions

a posteriori et les distributions de référence ou entre les distributions a posteriori obtenues à partir de différentes hypothèses a priori.

- **Le critère d'information d'Akaike et le critère d'information bayésien** :

Le critère d'information d'Akaike *(AIC)* et le critère d'information bayésien *(BIC)* sont deux mesures couramment utilisées pour évaluer la qualité des modèles statistiques, notamment dans le cadre des modèles bayésiens. Ces critères prennent en compte à la fois la qualité de l'ajustement du modèle et sa complexité, offrant ainsi un équilibre entre la précision et la simplicité.

La définition du critère d'information d'Akaike est donnée par :

$$AIC = 2k - 2ln(\hat{L})$$

où k est le nombre de paramètres dans le modèle et $\hat{L}$ est le maximum de la vraisemblance du modèle.

L'AIC évalue l'information perdue lors de l'utilisation d'un modèle pour représenter le processus qui a généré les données. Un modèle avec un AIC plus faible est généralement préféré. Il pénalise les modèles ayant un plus grand nombre de paramètres pour éviter le sur-ajustement.

La définition du critère d'information bayésien est donnée par :

$$BIC = ln(n)k - 2ln(\hat{L})$$

où n est le nombre d'observations, k est le nombre de paramètres dans le modèle, et $\hat{L}$ est le maximum de la vraisemblance du modèle.

Le BIC est similaire à l'AIC mais avec une pénalité plus forte pour les modèles avec un plus grand nombre de paramètres, en particulier lorsque le nombre d'observations est élevé. Il est particulièrement utile dans le cadre de la sélection de modèles dans une perspective bayésienne.

Les deux critères sont utilisés pour choisir entre plusieurs modèles concurrents. Le modèle avec le plus faible AIC ou BIC est généralement considéré comme le meilleur compromis entre la précision de l'ajustement et la complexité du modèle. Ces critères sont particulièrement utiles dans les situations où plusieurs modèles sont plausibles et où l'on cherche à identifier celui qui offre le meilleur équilibre entre simplicité et capacité à expliquer les données. Ils sont utilisés non seulement pour comparer des modèles avec des nombres de paramètres différents mais aussi pour évaluer l'efficacité de différents types de modèles, y compris dans les analyses bayésiennes.

En résumé, l'AIC et le BIC sont des outils précieux pour l'évaluation des modèles bayésiens, offrant une méthode pour juger la qualité des modèles en tenant compte de leur complexité et du nombre de paramètres. Ils aident à éviter le sur-ajustement tout en sélectionnant un modèle qui s'adapte bien aux données observées.

- **Analyse de sensibilité** :

L'analyse de sensibilité dans les modèles bayésiens examine comment la variation des distributions a priori affecte les résultats a posteriori. Elle permet de déterminer la robustesse des conclusions bayésiennes face aux incertitudes ou aux variations dans les choix des paramètres a priori. L'objectif est de mesurer l'influence des diverses hypothèses a priori sur les conclusions tirées du modèle bayésien.

Cette démarche vise à déterminer l'étendue de la sensibilité des résultats aux choix spécifiques des distributions a priori et à identifier celles qui exercent une influence notable sur les déductions finales du modèle.

L'analyse implique généralement de calculer la distribution a posteriori pour différents ensembles de distributions a priori. On peut définir un ensemble de distributions a priori $\Pi = \{\pi_1, \pi_2, ..., \pi_n\}$ où chaque π_i est une distribution a priori différente. Pour chaque distribution a priori π_i, on calcule la distribution a posteriori correspondante $P(\theta|\pi_i, X)$, où X est l'ensemble de données observées et θ représente les paramètres du modèle. On compare ensuite ces distributions a posteriori pour évaluer comment elles varient avec les différents choix de π_i. L'application d'une analyse de sensibilité peut être résumée de la manière suivante :

- **Sélectionner un ensemble de lois a priori** : Choisir un éventail de distributions a priori, allant des plus informatives aux plus non informatives ou vagues.
- **Calculer les distributions a posteriori** : Pour chaque loi a priori choisie, effectuer l'inférence bayésienne pour obtenir la distribution a posteriori correspondante.
- **Comparer les résultats** : Analyser les différences entre les distributions a posteriori obtenues à partir des différentes lois a priori. Cela peut impliquer l'utilisation de mesures quantitatives, telles que la divergence de Kullback-Leibler, ou des analyses qualitatives, telles que l'examen des changements dans les intervalles de crédibilité.
- **Interpréter les résultats** : Identifier les a priori qui ont un impact significatif sur les résultats et déterminer la robustesse des conclusions du modèle face aux variations des hypothèses a priori.

En conclusion, l'analyse de sensibilité est un outil crucial dans les modèles bayésiens pour comprendre l'impact des choix a priori sur les inférences. Elle aide à assurer que les conclusions du modèle sont fiables et ne dépendent pas de manière critique des hypothèses a priori spécifiques choisies.

- **Validation croisée** :

La validation croisée est une technique statistique essentielle pour évaluer la performance des modèles prédictifs, y compris dans le cadre bayésien. Elle permet de tester l'efficacité du modèle en utilisant des sous-ensembles de données pour entraîner et tester le modèle, fournissant ainsi une évaluation de la capacité prédictive du modèle.

L'objectif principal de la validation croisée est d'évaluer de façon fiable et précise les performances du modèle sur des données inédites, ce qui permet d'estimer sa capacité à généraliser et à fournir des prédictions exactes à partir de jeux de données limités ou restreints.

La méthode de validation croisée la plus couramment employée est la technique de la validation croisée à k blocs. Les données sont divisées en k sous-ensembles (ou blocs). Le modèle est formé en utilisant $k-1$ sous-ensembles et est ensuite évalué sur le sous-ensemble restant. Ce processus est répété k fois, avec chaque sous-ensemble utilisé exactement une fois comme ensemble de test. L'application de la validation croisée dans les modèles bayésiens peut être synthétisée comme suit :

- **Choix des distributions a priori** : Sélectionner différentes distributions a priori pour l'inférence bayésienne.
- **Division des données** : Diviser l'ensemble de données en sous-ensembles pour la formation

et le test.

- **Entraînement et test** : Pour chaque loi a priori sélectionnée, effectuer l'inférence bayésienne sur l'ensemble d'entraînement, puis utiliser le modèle résultant pour faire des prédictions sur l'ensemble de test.

- **Évaluation** : Utiliser des mesures de performance telles que l'erreur quadratique moyenne *(MSE)*, l'erreur absolue moyenne *(MAE)* ou d'autres métriques appropriées pour évaluer la qualité des prédictions.

- **Comparaison** : Comparer les performances obtenues avec différentes lois a priori pour évaluer leur impact sur la capacité prédictive du modèle.

La validation croisée dans les modèles bayésiens offre un moyen robuste d'évaluer la performance prédictive du modèle tout en tenant compte de l'incertitude et de la variabilité des données. Elle est essentielle pour s'assurer que les conclusions tirées du modèle sont fiables et transférables à de nouvelles données.

Conclusion

Ce mémoire a cherché à offrir une compréhension approfondie des dynamiques actuelles en assurance santé collective, avec un accent particulier sur l'adoption de l'approche bayésienne pour la tarification a posteriori. Nous avons exploré les principes fondamentaux de la tarification, différenciant les méthodes a priori et a posteriori, et soulignant l'importance d'adapter ces méthodes aux spécificités du secteur de l'assurance santé. Cette exploration théorique a servi de base à notre étude sur la façon dont les méthodes bayésiennes peuvent être utilisées pour affiner les modèles de tarification, en capturant efficacement la complexité et la variabilité des données du secteur.

Notre travail a également inclus la présentation et l'application pratique des modèles bayésiens, illustrés par des exemples concrets et appliqués à un échantillon de 30 entreprises. Ces analyses ont révélé l'efficacité de l'approche bayésienne, en particulier pour la tarification a posteriori dans le secteur de l'assurance santé collective. Cependant, notre étude a été limitée par la complexité algorithmique inhérente aux modèles bayésiens, en particulier les méthodes de Monte Carlo par chaînes de Markov, ce qui nous a empêchés d'appliquer ces méthodes à toutes les entreprises initialement envisagées dans le périmètre de notre étude. Pour surmonter ces défis algorithmiques et rendre l'application des modèles bayésiens plus pratique au sein des entreprises d'assurance, il est crucial de bien choisir la granularité de leur mise en œuvre. Il pourrait être judicieux d'appliquer ces modèles de manière ciblée, par exemple, à des produits spécifiques plutôt qu'à l'ensemble de l'entreprise, ou à des segments d'entreprise présentant des comptes déficitaires.

Les modèles hiérarchiques bayésiens ont démontré leur capacité à intégrer les caractéristiques de chaque entreprise, telles que les variables traditionnellement utilisées pour la tarification a priori. Cette approche permet d'incorporer des connaissances préalables détaillées dans le processus de modélisation, offrant ainsi une personnalisation et une adaptabilité accrues des tarifs. En outre, l'application de ces modèles nous offre la possibilité de vérifier et de valider nos hypothèses tarifaires de manière rigoureuse. Elle nous apporte une compréhension approfondie de l'impact des critères spécifiques à chaque entreprise sur la consommation des soins de santé, améliorant significativement notre capacité à interpréter et à ajuster les tarifs en fonction des caractéristiques et des risques associés à chaque profil d'entreprise.

L'étude a également démontré l'importance de la sélection des distributions a priori, cruciale pour la précision des estimations et l'alignement de ces dernières avec nos prévisions et attentes. Il a été observé que les fréquences et les montants de remboursement varient largement, nécessitant ainsi une modélisation complexe et sensible à ces variations. La granularité avec laquelle les modèles bayésiens sont appliqués s'est révélée être un facteur clé. Notre approche détaillée par regroupement d'actes et par entreprise a permis une compréhension plus fine des mécanismes des modèles bayésiens, malgré la possibilité d'une simplification en consolidant les divers actes en catégories plus larges.

En outre, nous avons abordé l'impact des variations temporelles sur l'estimation bayésienne, en soulignant la possibilité d'intégrer des méthodes avancées comme les processus gaussiens pour une analyse plus poussée. Par ailleurs, nous avons présenté des méthodes et des métriques pour l'évaluation et la comparaison de différents modèles. Cette démarche se révèle particulièrement pertinente lors de l'expérimentation avec plusieurs distributions a priori variées. Ces outils offrent une perspective cruciale, non seulement pour l'optimisation des modèles en cours mais aussi pour l'exploration de

nouvelles approches, renforçant ainsi notre capacité à adapter et affiner les méthodes bayésiennes selon les besoins spécifiques de l'assurance santé collective.

En définitive, cette étude souligne l'importance et l'utilité de l'approche bayésienne et des modèles hiérarchiques dans le domaine de l'actuariat moderne, mettant en avant plusieurs avantages significatifs. Tout d'abord, l'intégration des connaissances et de l'expertise préalables à travers les distributions a priori enrichit les estimations dès le départ. Ensuite, cette approche permet une mise à jour dynamique des estimations tarifaires, adaptant continuellement les calculs à l'arrivée de nouvelles données. Par ailleurs, elle fournit une mesure explicite de l'incertitude, ce qui est crucial pour une meilleure gestion des risques et une planification stratégique plus efficace. Enfin, l'approche bayésienne facilite une personnalisation avancée des tarifs, permettant une tarification ajustée aux spécificités de chaque entreprise ou groupe d'assurés grâce à une analyse détaillée.

Cependant, cette méthode présente aussi des défis. La complexité algorithmique des modèles peut exiger des ressources de calcul importantes, particulièrement avec de grands volumes de données. L'efficacité de l'approche repose fortement sur la qualité et la granularité des données disponibles, limitant son application en l'absence de données fiables. La sélection des distributions a priori nécessite une expertise approfondie et peut introduire des biais si mal gérée. Enfin, la complexité des modèles bayésiens peut rendre leurs résultats difficiles à interpréter pour les non-spécialistes, soulignant la nécessité d'une communication claire et d'une formation adéquate.

Ces avantages et inconvénients soulignent l'importance d'une mise en œuvre réfléchie et bien gérée de l'approche bayésienne en actuariat. Malgré les défis, les perspectives prometteuses offertes par ces méthodes, des tarifications plus équitables, adaptatives et transparentes, représentent un pas significatif vers une compréhension plus précise et une gestion plus fine des risques dans le secteur de l'assurance santé collective. L'alignement entre précision analytique et perspicacité commerciale offre aux assureurs les outils pour ajuster les tarifs avec une granularité qui reflète fidèlement le risque inhérent à chaque groupe spécifique, pavant ainsi le chemin vers des méthodes de tarification et de gestion des risques plus avancées et intuitives dans le monde de l'assurance.

Bibliographie

[1] Ministère des Solidarités et des Familles. La sécurité sociale : fonctionnement, branches et caisses. 2019.

[2] La Sécurité Sociale. Les chiffres clés de la Sécurité Sociale 2022.

[3] DREES. Rapport 2022 - Sur la situation financière des organismes complémentaires assurant une couverture santé. 2022.

[4] L'Assemblée nationale et le Sénat. Loi n° 2013-504 du 14 juin 2013. Legifrance, 2013.

[5] Ministère de la santé et de la prévention. 100% SANTÉ - Des soins pour tous, 100% pris en charge. 2018.

[6] SGAM Malakoff Humanis. Rapport sur la Solvabilité et la Situation Financière 2022.

[7] Hans Bühlmann and Alois Gisler. *A course in credibility theory and its applications*. Springer, 2005.

[8] BOCQUAIRE Edith, Nadine CHARLES, and Roger MILLOT. *Pratique de l'assurance santé*. L'Argus de l'assurance, 2017.

[9] Cameron Davidson-Pilon. *Bayesian methods for hackers : probabilistic programming and Bayesian inference*. Addison-Wesley Professional, 2015.

[10] Osvaldo Martin. *Bayesian analysis with python*. Packt Publishing Ltd, 2016.

[11] Osvaldo A Martin, Ravin Kumar, and Junpeng Lao. *Bayesian modeling and computation in python*. CRC Press, 2021.