

Notes de cours sur la théorie de la crédibilité

M1 actuariat

Pierre Ailliot

Année 2023-2024

1 Introduction

La théorie de la crédibilité est un ensemble de techniques utilisées par les actuaires pour faire de la tarification en présence d'un portefeuille hétérogène. Un portefeuille hétérogène est un ensemble de contrats (ou "assurés") avec des profils de risque différents. Il semble alors inéquitable de réclamer la même prime d'assurance à tous les assurés. La théorie de la crédibilité va permettre de faire une tarification dans laquelle les contrats qui présentent un profil de risque plus dangereux vont payer une prime plus importante que les contrats avec un profil de risque moins dangereux.

Données. La théorie de la crédibilité effectue une tarification basée sur l'expérience, c'est à dire basée sur un historique des sinistres passés pour chaque contrat assuré. Il faut donc disposer d'un historique assez conséquent (au moins 5 ans). Ces méthodes sont par exemple utilisés en assurance automobile (parc automobile des entreprises par exemple) et pour la prévoyance collective des entreprises ou de tels historiques existent. Par contre, elles ne peuvent pas être utilisées en assurance vie (on ne meurt qu'une fois) ou en assurance habitation (trop peu de sinistres). En pratique, on dispose d'un jeu de données ("portefeuille") de la forme du tableau 1.

Chaque ligne représente un contrat et chaque colonne représente une année. On notera $x_{i,j}$ le montant (ou éventuellement le nombre) de sinistres pour le contrat $i \in \{1, \dots, m\}$ l'année $j \in \{1, \dots, n\}$. m désigne le nombre de contrats dans le portefeuille et n le nombre d'années d'observation.

Exemple. On considère un portefeuille de 10 contrats. Les montants annuels des sinistres observés au cours de 10 années sont donnés dans le tableau 1.

Année	1	2	3	4	5	6	7	8	9	10	Moyenne
Contrat 1	0	0	10	0	0	0	0	5	0	0	1.5
Contrat 2	15	0	0	12	0	0	5	10	0	0	4.2
Contrat 3	0	0	2	0	0	0	3	0	5	0	1
Contrat 4	0	20	0	0	15	0	0	25	0	0	6
Contrat 5	0	0	0	0	5	0	0	10	0	0	1.5
Contrat 6	0	0	0	3	0	0	2	0	0	5	1
Contrat 7	5	0	0	0	2	0	5	0	0	0	1.2
Contrat 8	0	30	0	15	0	0	20	0	0	20	8.5
Contrat 9	0	0	2	0	0	0	5	0	2	0	0.9
Contrat 10	0	0	0	0	0	0	0	0	10	0	1

Table 1: Sinistralité observée pour 10 contrats pendant 10 ans.

En pratique, l'objectif de la théorie de la crédibilité est de prédire le montant des sinistres l'année 11 pour chaque contrat, c'est à dire de 'compléter' le tableau 1 en faisant la meilleure prédiction possible de la 11ème colonne.

Formalisation statistique. On supposera que les observations $x = (x_{i,j})_{i \in \{1, \dots, m\}, j \in \{1, \dots, n\}}$ sont une réalisation d'un vecteur aléatoire $X = (X_{i,j})_{i \in \{1, \dots, m\}, j \in \{1, \dots, n\}}$. La théorie de la crédibilité a pour objectif principal d'estimer la **PRIME PURE** des différents contrats qui est définie comme l'espérance du montant des sinistres pour les différents contrats l'année $n + 1$ connaissant la sinistralité passée, c'est à dire

$$\mu_{i,n+1} = E[X_{i,n+1} | X = x].$$

On rappelle que cette espérance conditionnelle peut s'interpréter comme la meilleur prédiction (au sens des moindres carrés) des sinistres l'année $n + 1$ connaissant les sinistres passés de l'assuré. Le calcul de cette espérance conditionnelle va passer par une modélisation de la loi jointe du vecteur aléatoire X qui décrit le portefeuille.

Dans la suite du cours, on supposera que les différents contrats sont indépendants. Dans ce cas, la prime pure du contrat i s'écrit

$$\mu_{i,n+1} = E[X_{i,n+1} | X_{i,1} = x_{i,1}, \dots, X_{i,n} = x_{i,n}].$$

Deux stratégies opposées sont possibles pour estimer la prime pure d'un contrat.

1. On suppose que le portefeuille est homogène. Les risques sont alors les mêmes pour tous les contrats l'année $n + 1$, c'est à dire

$$\mu = \mu_{i,n+1}$$

pour $i \in \{1, \dots, m\}$ avec $\mu \in \mathbb{R}$ la prime collective. Tous les assurés paieront alors la même prime l'année $n + 1$. Un estimateur naturel de cette prime collective est la **prime collective empirique**

$$\bar{X} = \frac{1}{nm} \sum_{i=1}^m \sum_{j=1}^n X_{i,j}$$

c'est à dire la moyenne des sinistres sur tous les contrats et les années d'observation.

Sur l'exemple du tableau 1, on observe une forte variabilité entre les contrats et on peut avoir des doutes sur l'hypothèse d'homogénéité (par exemple les contrats 2, 4 et 8 semblent avoir un profil de risque plus dangereux que les autres contrats). On peut vérifier l'hypothèse d'homogénéité du portefeuille en réalisant un test d'hypothèses pour comparer les sinistres moyens des différents contrats. Plus précisément, supposons que les sinistres $(X_{i,j})_{i \in \{1, \dots, m\}, j \in \{1, \dots, n\}}$ sont indépendants et que $X_{i,j} \sim \mathcal{N}(\mu_i, \sigma^2)$ avec μ_i l'espérance du montant des sinistres du contrat i . On peut tester (test d'analyse de la variance à un facteur, cf cours sur le modèle linéaire) :

$$H_0 : \mu_1 = \dots = \mu_m \text{ contre } H_1 : H_0 \text{ fausse.}$$

La p-value du test est $p_v = .025$ (cf exercice 5) et on refuse donc H_0 avec un risque de première espèce $\alpha = 5\%$.

En présence d'un portefeuille hétérogène, il peut paraître inéquitable de faire payer la même prime à tous les assurés. En effet, les assurés avec un profil de risque faible vont payer un surcroît de prime pour dédommager les sinistres causés par les individus les plus risqués. L'assureur risque en particulier de perdre les "bons" assurés qui auront le sentiment de trop payer (on parle de risque d'antisélection en assurance).

2. À l'opposé, on peut individualiser complètement la prime en se basant sur l'historique de chaque contrat. La prime pour le contrat i l'année $n + 1$ sera alors estimée par la **prime individuelle empirique**

$$\bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{i,j}$$

c'est à dire que la prime de risque d'un assuré est la moyenne des sinistres de cet assuré sur les années précédentes. Cette tarification serait opposée au principe même de l'assurance puisque le risque n'est plus mutualisé. Imaginez par exemple un assuré qui n'a aucun sinistre les premières années : il sera couvert par l'assurance sans payer de prime d'assurance. Si il a ensuite un sinistre important au bout de quelques années, sa prime d'assurance augmentera subitement.

Face à un portefeuille hétérogène, une première stratégie pour l'assureur consiste à segmenter le portefeuille en plus petits portefeuilles homogènes et à appliquer la tarification collective à chacun des groupes créés. Il peut cependant être difficile d'obtenir des groupes homogènes en terme de profils de risque.

Forme générale d'une prime de crédibilité. La théorie de la crédibilité cherche un compromis entre les deux stratégies de tarification opposées décrites ci-dessus. On utilise alors une prime pour l'année $n + 1$ de la forme

$$p_{i,n+1} = \lambda \bar{X}_i + (1 - \lambda) \bar{X}$$

avec $\lambda \in [0, 1]$ le **facteur de crédibilité** qui mesure la confiance accordée à (ou "crédibilité" de) la prime individuelle $\bar{X}_i$. On retrouve les cas extrêmes

- si $\lambda = 1$, alors $p_{i,n+1} = \bar{X}_i$ (tarification individuelle sans mutualisation),
- si $\lambda = 0$, alors $p_{i,n+1} = \bar{X}$ (tarification collective sans individualisation).

Equilibre financier. Lorsqu'il fait la tarification d'un portefeuille, l'assureur cherche

1. à charger assez de primes pour couvrir les sinistres (équilibre financier),
2. à répartir équitablement ces primes entre les assurés.

On peut remarquer que la valeur de λ ne modifie pas l'équilibre financier de l'assureur. En effet, la prime totale collectée par l'assureur pour l'année $n + 1$ est égale

$$p_{n+1} = \sum_{i=1}^m p_{i,n+1} = \lambda \sum_{i=1}^m \bar{X}_i + (1 - \lambda) \sum_{i=1}^m \bar{X} = m\bar{X}$$

puisque $\sum_{i=1}^m \bar{X}_i = m\bar{X}$. La prime totale collectée par l'assureur ne dépend donc pas de λ . Le choix de λ va donc être fait de telle manière à garantir l'équité entre les assurés.

Il existe deux grandes familles de méthodes en théorie de la crédibilité : la crédibilité bayésienne et la crédibilité linéaire. La crédibilité bayésienne sera abordée très rapidement à partir d'un exercice au chapitre 2. La crédibilité linéaire sera détaillée dans le chapitre 3.

2 Crédibilité bayésienne

Exercice 1 *Un portefeuille d'assurance automobile est composé avec une proportion π de bons conducteurs et de $1 - \pi$ de mauvais conducteurs. Les bons conducteurs ont une probabilité p_1 d'avoir un accident, alors que la probabilité est p_2 pour les mauvais conducteurs. On suppose que le coût d'un accident est de 1000 et qu'au plus un accident peut survenir dans une année. On prendra les valeurs numériques suivantes pour les applications : $\pi = 75\%$, $p_1 = 1/15$, $p_2 = 1/10$.*

1. *On choisit un assuré au hasard (on ne sait si il s'agit d'un bon ou d'un mauvais conducteur). Quelle est la probabilité que cet assuré ait un accident la première année? Quelle est la prime pure de cet assuré la première année? On répondra à cette question en essayant de formaliser le problème d'un point de vue probabiliste. On notera*

- Θ_i la variable aléatoire à valeurs dans $\{B, M\}$ (B pour Bon et M pour Mauvais) qui décrit le groupe de l'assuré numéro i qui a été choisi au hasard,
- $N_{i,j}$ la variable aléatoire à valeurs dans $\{0, 1\}$ qui décrit la survenance d'un sinistre l'année j pour l'assuré i ,
- $X_{i,j}$ la variable aléatoire qui décrit le montant des sinistres l'année j pour l'assuré i .

On explicitera la loi des ces variables aléatoires puis on calculera les quantités demandées.

2. L'assuré choisi dans la question précédente a eu un accident lors de la première année. On explicitera les hypothèses probabilistes utilisées pour faire les calculs ci-dessous.

- Quelle est la probabilité qu'il s'agisse d'un bon conducteur?
- Quelle est la prime pure de cet assuré pour la seconde année?

Notons

- Θ_i la variable aléatoire (v.a.) qui décrit le groupe de l'assuré. Θ_i est une v.a. à valeurs dans $\{B, M\}$ (B pour Bon et M pour Mauvais) avec $P(\Theta_i = B) = \pi$.
- $N_{i,j}$ la variable aléatoire qui décrit le nombre d'accident pour l'assuré i l'année j . D'après l'énoncé $N_{i,j}$ est une v.a. à valeurs dans $\{0, 1\}$ avec $P(N_{i,j} = 1 | \Theta_i = B) = p_1$ et $P(N_{i,j} = 1 | \Theta_i = M) = p_2$.
- $X_{i,j}$ la variable aléatoire qui décrit le montant des sinistres pour l'assuré i l'année j . D'après l'énoncé, on a $X_{i,j} = 1000N_{i,j}$.

1. La probabilité que l'assuré i choisi au hasard ait un accident la première année est, par la *formule des probabilités totales*,

$$P(N_{i,1} = 1) = P(N_{i,1} = 1 | \Theta_i = B)P(\Theta_i = B) + P(N_{i,1} = 1 | \Theta_i = M)P(\Theta_i = M) = p_1\pi + p_2(1 - \pi).$$

L'application numérique donne $E[N_{i,1}] = P(N_{i,1} = 1) = 0.075$. Par définition, la prime pure la première année est donnée par

$$E[X_{i,1}] = E[1000N_{i,1}].$$

L'application numérique donne une prime pure de 75.

2. (a) D'après la *formule de Bayes*

$$P(\Theta_i = B | N_{i,1} = 1) = \frac{P(N_{i,1} = 1 | \Theta_i = B)P(\Theta_i = B)}{P(N_{i,1} = 1)} = \frac{p_1\pi}{p_1\pi + p_2(1 - \pi)}$$

L'application numérique donne $P(\Theta_i = B | N_{i,1} = 1) = 2/3$.

(b) Par définition, si on suppose que les contrats sont indépendants, la prime pure est donnée par $E[X_{i,2} | X_{i,1} = 1000] = 1000P(N_{i,2} = 1 | N_{i,1} = 1)$.

Par la formule des probabilités totales, on a

$$P(N_{i,2} = 1 | N_{i,1} = 1) = P(N_{i,2} = 1, \Theta_i = B | N_{i,1} = 1) + P(N_{i,2} = 1, \Theta_i = M | N_{i,1} = 1)$$

puis en utilisant la formule de Bayes

$$\begin{aligned} P(N_{i,2} = 1 | N_{i,1} = 1) = \\ \frac{P(N_{i,2} = 1, N_{i,1} = 1 | \Theta_i = B)P(\Theta_i = B) + P(N_{i,2} = 1, N_{i,1} = 1 | \Theta_i = M)P(\Theta_i = M)}{P(N_{i,1} = 1)}. \end{aligned}$$

Afin de pouvoir calculer les quantités ci-dessus à partir des quantités données dans l'énoncé, il faut faire des hypothèses supplémentaires. En crédibilité, on suppose généralement que les sinistres d'un assuré les années successives sont **conditionnellement indépendantes** connaissant le "profil de risque" Θ_i , c'est à dire que $P(N_{i,2} = 1, N_{i,1} = 1 | \Theta_i = B) = P(N_{i,2} = 1 | \Theta_i = B)P(N_{i,1} = 1 | \Theta_i = B)$ et $P(N_{i,2} = 1, N_{i,1} = 1 | \Theta_i = M) = P(N_{i,2} = 1 | \Theta_i = M)P(N_{i,1} = 1 | \Theta_i = M)$. Avec cette hypothèse et après simplification, on obtient que

$$\begin{aligned} P(N_{i,2} = 1 | N_{i,1} = 1) \\ &= P(N_{i,2} = 1 | \Theta_i = B)P(\Theta_i = B | N_{i,1} = 1) + P(N_{i,2} = 1 | \Theta_i = M)P(\Theta_i = M | N_{i,1} = 1) \\ &= p_1 \frac{p_1 \pi}{p_1 \pi + p_2(1 - \pi)} + p_2 \frac{p_2(1 - \pi)}{p_1 \pi + p_2(1 - \pi)}. \end{aligned}$$

NB. Cette relation, qui peut sembler intuitive, n'est pas toujours vraie. Nous avons du supposer l'indépendance conditionnelle pour la démontrer.

L'application numérique donne $P(N_{i,2} = 1 | N_{i,1} = 1) = 0.0778$ puis la prime pure pour la deuxième année $E(X_2 | N_{i,1} = 1) = 1000P(N_{i,2} = 1 | N_{i,1} = 1) = 77.8$.

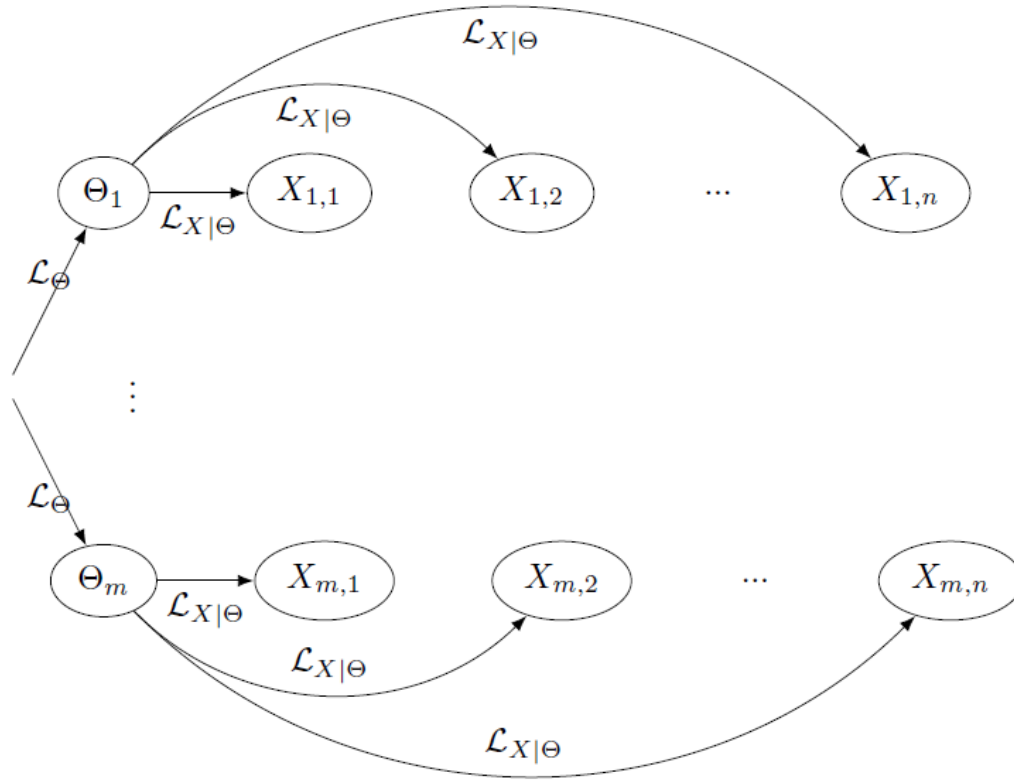


Figure 1: Diagramme d'indépendance conditionnelle en crédibilité bayésienne

Extension. L'exercice ci-dessus donne un exemple de modèle de crédibilité bayésienne. Les modèles utilisés en théorie de la crédibilité ont généralement la même structure, avec

- une variable Θ_i (non observée) qui décrit le profil de risque des assurés,
- l'hypothèse d'indépendance conditionnelle pour les sinistres conditionnellement au profil de risque.

Ces hypothèses sont résumées sur la diagramme d'indépendance conditionnelle de la Figure 1.

Dans l'exercice, la loi de Θ_i et la loi conditionnelle des sinistres sont binaires (respectivement bon/mauvais et sinistre/pas sinistre, avec le même montant pour tous les sinistres). Les modèles utilisés en crédibilité bayésienne utilisent généralement des lois continues. Le modèle le plus classique est le **modèle Exponentielle-Gamma** dans lequel on suppose que

- les profils de risque suivent une loi Gamma,
- conditionnellement aux profils de risque, les sinistres suivent une loi Exponentielle dont le paramètre dépend du profil de risque.

Avec ces hypothèses, on peut calculer la prime pure et montrer qu'elle s'écrit comme une prime de crédibilité, c'est à dire que

$$E[X_{i,n+1}|X_{i,1} = x_{i,1}, \dots, X_{i,n} = x_{i,n}] = \lambda \bar{X}_i + (1 - \lambda) \bar{X}$$

avec une expression explicite pour λ .

Une limitation des méthodes de crédibilité bayésienne est qu'elles nécessitent de faire des hypothèses paramétriques qui peuvent sembler parfois arbitraires et difficiles à valider en pratique. A l'opposé, les méthodes de crédibilité linéaire discutées dans le chapitre suivent ne sont pas basées sur des hypothèses paramétriques.

3 Crédibilité linéaire

De manière générale, la crédibilité linéaire s'intéresse aux primes qui s'écrivent comme une combinaison linéaire des sinistres passés de l'assuré, c'est à dire de la forme

$$p_{i,n+1} = c_0 + c_1 X_{i,1} + \dots + c_n X_{i,n}.$$

En général, la prime de crédibilité linéaire $p_{i,n+1}$ sera différente de la prime pure $\mu_{i,n+1} = E[X_{i,n+1}|X_{i,1}, \dots, X_{i,n}]$ si celle-ci ne s'écrit pas comme une fonction linéaire de $X_{i,1}, \dots, X_{i,n}$. L'idée est de chercher la meilleure approximation de $\mu_{i,n+1}$ qui s'écrit comme une fonction linéaire de $X_{i,1}, \dots, X_{i,n}$. Les coefficients $(c_0, c_1, \dots, c_n)$ s'interprètent comme des poids donnés aux sinistres passés pour prévoir "au mieux" les sinistres l'année $n + 1$. Ces poids étant inconnus, on doit les estimer à partir des observations disponibles. Comme pour le modèle de régression linéaire, ils sont estimés par la méthode des moindres carrés en minimisant la fonction objectif

$$F(\lambda_0, \lambda_1, \dots, \lambda_n) = E[(X_{i,n+1} - (\lambda_0 + \lambda_1 X_{i,1} + \dots + \lambda_n X_{i,n}))^2] \quad (1)$$

qui décrit l'écart quadratique entre la prime $p_{i,n+1}$ et les sinistres $X_{i,n+1}$ l'année $n + 1$. On pose ensuite

$$(c_0, c_1, \dots, c_n) = \operatorname{argmin}_{(\lambda_0, \lambda_1, \dots, \lambda_n) \in \mathbb{R}^{n+1}} F(\lambda). \quad (2)$$

Quand on développe l'expression (1), on voit apparaître des termes de la forme $E[X_{i,j} X_{i,j'}] = \operatorname{cov}(X_{i,j}, X_{i,j'}) + E[X_{i,j}]E[X_{i,j'}]$ et on doit donc connaître ces quantités pour le calcul et la minimisation de la fonction objectif F . Pour cela, la crédibilité linéaire fait des hypothèses sur la loi du couple de variables aléatoires $(X_{i,j}, X_{i,j'})$ (cf paragraphe 3.1). La formule de la variance totale permet ensuite de déduire des expressions pour $\operatorname{cov}(X_{i,j}, X_{i,j'})$ (cf paragraphe 3.2). La minimisation de F est discutée en 3.3 et les aspects pratiques en 3.4.

3.1 Hypothèses du modèle de Bühlmann (1967)

Dans le modèle de Bühlmann, on suppose que chaque contrat a un profil de risque et on notera Θ_i la variable aléatoire qui décrit le profil de risque du contrat $i \in \{1, \dots, m\}$. La structure du portefeuille est donc caractérisée par les variables aléatoires qui apparaissent dans le tableau 2.

	Profils de risque	Sinistres année 1	Sinistres année 2	...	Sinistres année n
Contrat 1	Θ_1	$X_{1,1}$	$X_{1,2}$	...	$X_{1,n}$
Contrat 2	Θ_2	$X_{2,1}$	$X_{2,2}$	...	$X_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
Contrat m	Θ_m	$X_{m,1}$	$X_{m,2}$	...	$X_{m,n}$

Table 2: Structure du portefeuille dans le modèle de Bühlman.

En pratique, le profil de risque Θ_i n'est pas observé (on parle de variable latente en statistique). On doit donc pouvoir estimer la prime sans utiliser cette variable.

Dans le modèle de Bühlmann, différentes hypothèses sont faites sur la loi des variables aléatoires qui apparaissent dans le tableau 2. Ces hypothèses utilisent la covariance conditionnelle définie ci-dessous.

Définition 1 La covariance conditionnelle des sinistres $X_{i,j}$ et $X_{i,j'}$ pour les années j et j' connaissant le profil de risque Θ_i est définie par

$$\text{cov}(X_{i,j}, X_{i,j'} | \Theta_i) = E[(X_{i,j} - E[X_{i,j} | \Theta_i]) (X_{i,j'} - E[X_{i,j'} | \Theta_i]) | \Theta_i].$$

Si $j = j'$ on obtient la variance conditionnelle

$$\text{var}(X_{i,j} | \Theta_i) = \text{cov}(X_{i,j}, X_{i,j} | \Theta_i) = E[(X_{i,j} - E[X_{i,j} | \Theta_i])^2 | \Theta_i].$$

Remarque 1 Les définitions ci-dessus sont obtenues en remplaçant les espérances par les espérances conditionnelles dans les définitions usuelles de la variance et de la covariance.

Le modèle de Bühlmann fait les hypothèses suivantes sur la structure du portefeuille.

Hypothèses de Bühlmann

1. Les profils de risque Θ_i pour $i \in \{1, \dots, n\}$ sont des variables aléatoires iid.
2. Les contrats, définis pour $i \in \{1, \dots, n\}$ par le vecteur aléatoire $C_i = (\Theta_i, X_{i,1}, \dots, X_{i,n})$, sont des vecteurs aléatoires indépendants.
3. $E[X_{i,j} | \Theta_i] = m(\Theta_i)$ et $\text{var}(X_{i,j} | \Theta_i) = v(\Theta_i)$ avec m et v des fonctions (mesurables) qui ne dépendent ni de i (hypothèse d'homogénéité "inter-contrats") ni de j (hypothèse de stationnarité temporelle ou d'homogénéité "intra-contrats").
4. $\text{cov}(X_{i,j}, X_{i,j'} | \Theta_i) = 0$ si $j \neq j'$ (les sinistres d'un assuré pour les années j et j' sont non corrélés conditionnellement au profil de risque). Nous verrons dans la suite que cette hypothèse n'implique pas que $\text{cov}(X_{i,j}, X_{i,j'}) = 0$ (la corrélation entre $X_{i,j}$ et $X_{i,j'}$ est créée par le profil de risque).

Les hypothèses (3-4) sont vraies en particulier lorsque les sinistres d'un contrat satisfont les hypothèses de la Figure 1, c'est à dire sont conditionnellement indépendants connaissant le profil de risque

$$p(X_{i,1} = x_{i,1}, \dots, X_{i,n} = x_{i,n} | \Theta_i = \theta_i) = \prod_{j=1}^n p(X_{i,j} = x_{i,j} | \Theta_i = \theta_i)$$

et que la loi conditionnelle des sinistres $\mathcal{L}_{X_{i,j}|\Theta=\theta_i}$ ne dépend ni du contrat i (homogénéité inter-contrat) ni de l'année j (homogénéité intra-contrat). Avec ces hypothèses (légèrement plus fortes que celle du modèle de Bühlmann), on peut facilement simuler un portefeuille qui satisfait les hypothèses du modèle de Bühlmann avec le pseudo-code ci-dessous :

- Simuler un échantillon iid $(\theta_1, \dots, \theta_m)$ selon une loi $\mathcal{L}_\theta$
- Pour $i \in \{1, \dots, n\}$, simuler un échantillon iid $(x_{i,1}, \dots, x_{i,n})$ selon la loi conditionnelle $\mathcal{L}_{X|\Theta=\theta_i}$.

Sous les hypothèses (1-4) du modèle de Bühlmann, on peut vérifier que les quantités suivantes sont bien définies car elles ne dépendent pas de i et de j :

- $\mu = E[X_{i,j}]$ est appelée la **prime collective**.
- $\Sigma^2 = E[\text{var}(X_{i,j}|\Theta_i)] = E[(X_{i,j} - E[X_{i,j}|\Theta_i])^2]$ mesure la variabilité résiduelle au cours des années une fois qu'on a enlevé l'effet du profil de risque Θ_i (variance intra-contrat). En particulier, si $\Sigma^2 = 0$ alors $X_{i,j} = E[X_{i,j}|\Theta_i]$ presque sûrement (p.s.) et donc $X_{i,j}$ est une fonction déterministe de Θ_i . En pratique, cela signifie que les lignes du tableau de données $(x_{i,j})$ sont constantes (p.s.) et donc que les sinistres d'un contrat donné sont les mêmes tous les ans.
- $M^2 = \text{var}(E[X_{i,j}|\Theta_i])$ mesure l'hétérogénéité du portefeuille (variance inter-contrat). En particulier, si $M^2 = 0$ alors la variable aléatoire $E[X_{i,j}|\Theta_i]$ est une variable aléatoire constante (p.s.). Cela signifie que le profil de risque n'apporte pas d'information sur la sinistralité et que le portefeuille est homogène.

Exercice 2 . On suppose qu'un portefeuille d'assurance vérifie

$$X_{i,j} = f(\theta_i) + \epsilon_{i,j}$$

avec $f : \mathbb{R} \rightarrow \mathbb{R}$ une fonction mesurable, $(\theta_1, \dots, \theta_m)$ des variables aléatoires réelles identiquement distribuées et $(\epsilon_{i,j})_{i \in \{1, \dots, m\}, j \in \{1, \dots, n\}}$ des variables aléatoires réelles centrées identiquement distribuées. On supposera de plus que les variables aléatoires $(\theta_i, \epsilon_{i,j})_{i \in \{1, \dots, m\}, j \in \{1, \dots, n\}}$ sont indépendantes.

1. Montrer que le portefeuille satisfait les hypothèses du modèle de Bühlmann et exprimer les quantités μ , Σ^2 et M^2 .
2. Calculer $\text{cov}(X_{i,j}, X_{i',j'})$ pour $i, i' \in \{1, \dots, m\}$ et $j, j' \in \{1, \dots, n\}$

1. On vérifie facilement que les hypothèses (1-2) sont satisfaites. On a

- $E[X_{i,j}|\theta_i] = E[f(\theta_i) + \epsilon_{i,j}|\theta_i] = E[f(\theta_i)|\theta_i] = f(\theta_i)$ car $\epsilon_{i,j}$ est indépendant de θ_i , f est mesurable et $E[\epsilon_{i,j}] = 0$. Comme les variables aléatoires $(\theta_1, \dots, \theta_m)$ sont identiquement distribuées, on en déduit que les lois de $f(\theta_i)$ et donc de $E[X_{i,j}|\theta_i]$ ne dépendent ni de i ni de j .
- $\text{var}(X_{i,j}|\theta_i) = E[(X_{i,j} - E[X_{i,j}|\theta_i])^2|\theta_i] = E[\epsilon_{i,j}^2|\theta_i] = E[\epsilon_{i,j}^2] = \text{var}(\epsilon_{i,j})$ car $\epsilon_{i,j}$ est indépendant de θ_i . Comme les variables aléatoires $(\epsilon_{i,j})_{i \in \{1, \dots, m\}, j \in \{1, \dots, n\}}$ sont identiquement distribuées, on en déduit que la loi de $\text{var}(X_{i,j}|\theta_i)$ est une constante qui ne dépend ni de i ni de j . On a donc vérifié que l'hypothèse (3) est satisfaite.
-

$$\begin{aligned} \text{cov}(X_{i,j}, X_{i,j'}|\Theta_i) &= E[(X_{i,j} - E[X_{i,j}|\Theta_i])(X_{i,j'} - E[X_{i,j'}|\Theta_i])|\Theta_i] \\ &= E[(X_{i,j} - f(\theta_i))(X_{i,j'} - f(\theta_i))|\Theta_i] \\ &= E[\epsilon_{i,j}\epsilon_{i,j'}|\Theta_i]. \end{aligned}$$

On a supposé que les variables aléatoires $\epsilon_{i,j}$ et $\epsilon_{i,j'}$ sont indépendantes de θ_i . On en déduit que

$$\text{cov}(X_{i,j}, X_{i,j'} | \Theta_i) = E[\epsilon_{i,j}\epsilon_{i,j'} | \Theta_i] = E[\epsilon_{i,j}\epsilon_{i,j'}].$$

Par ailleurs, comme les variables aléatoires $\epsilon_{i,j}$ et $\epsilon_{i,j'}$ sont indépendantes si $j \neq j'$, on en déduit que

$$E[\epsilon_{i,j}\epsilon_{i,j'}] = E[\epsilon_{i,j}]E[\epsilon_{i,j'}] = 0$$

puis finalement que $\text{cov}(X_{i,j}, X_{i,j'} | \Theta_i) = 0$. On a donc vérifié que l'hypothèse (4) est satisfaite.

On a bien vérifié les hypothèses (1-4) du modèle de Bühlmann. On a $\mu = E[X_{i,j}] = E[f(\theta_i)]$, $\Sigma^2 = E[\text{var}(X_{i,j} | \Theta_i)] = \text{var}[\epsilon_{i,j}]$ et $M^2 = \text{var}(E[X_{i,j} | \Theta_i]) = \text{var}(f(\theta_i))$.

2. On a

$$\begin{aligned} \text{cov}(X_{i,j}, X_{i',j'}) &= \text{cov}(f(\theta_i) + \epsilon_{i,j}, f(\theta_{i'}) + \epsilon_{i',j'}) \\ &= \text{cov}(f(\theta_i), f(\theta_{i'})) + \text{cov}(\epsilon_{i,j}, \epsilon_{i',j'}) \\ &= \begin{cases} 0 & \text{si } i \neq i' \\ M^2 & \text{si } i = i' \text{ et } j \neq j' \\ M^2 + \Sigma^2 & \text{si } i = i' \text{ et } j = j' \end{cases} \end{aligned}$$

3.2 Formule de la variance totale

Proposition 1 Si $X_{i,j}$, $X_{i,j'}$ et Θ_i sont trois variables aléatoires, alors

$$\text{cov}(X_{i,j}, X_{i,j'}) = E(\text{cov}(X_{i,j}, X_{i,j'} | \Theta_i)) + \text{cov}(E[X_{i,j} | \Theta_i], E(X_{i,j'} | \Theta_i)).$$

Preuve 1 Par définition

$$\begin{aligned} \text{cov}(X_{i,j}, X_{i,j'}) &= E[(X_{i,j} - E[X_{i,j}]) (X_{i,j'} - E[X_{i,j'}])] \\ &= E[((X_{i,j} - E[X_{i,j} | \Theta_i]) + (E[X_{i,j} | \Theta_i] - E[X_{i,j}])) ((X_{i,j'} - E[X_{i,j'} | \Theta_i]) + (E[X_{i,j'} | \Theta_i] - E[X_{i,j'}]))] \\ &= E[(X_{i,j} - E[X_{i,j} | \Theta_i]) (X_{i,j'} - E[X_{i,j'} | \Theta_i])] \end{aligned} \tag{3}$$

$$+ E[(E[X_{i,j} | \Theta_i] - E[X_{i,j}]) (E[X_{i,j'} | \Theta_i] - E[X_{i,j'}])] \tag{4}$$

$$+ E[(X_{i,j} - E[X_{i,j} | \Theta_i]) (E[X_{i,j'} | \Theta_i] - E[X_{i,j'}])] \tag{5}$$

$$+ E[(E[X_{i,j} | \Theta_i] - E[X_{i,j}]) (X_{i,j'} - E[X_{i,j'} | \Theta_i])]. \tag{6}$$

Calculons maintenant les 4 termes qui apparaissent dans la dernière expression. Pour le premier terme (3), en utilisant que $E[Y] = E[E[Y | \theta]]$, on obtient

$$\begin{aligned} &E[(X_{i,j} - E[X_{i,j} | \Theta_i]) (X_{i,j'} - E[X_{i,j'} | \Theta_i])] \\ &= E[E[(X_{i,j} - E[X_{i,j} | \Theta_i]) (X_{i,j'} - E[X_{i,j'} | \Theta_i]) | \Theta_i]] \\ &= E(\text{cov}(X_{i,j}, X_{i,j'} | \Theta_i)). \end{aligned}$$

Pour le deuxième terme (4), comme $E[E[X_{i,j} | \Theta_i]] = E[X_{i,j}]$, on obtient

$$\begin{aligned} &E[(E[X_{i,j} | \Theta_i] - E[X_{i,j}]) (E[X_{i,j'} | \Theta_i] - E[X_{i,j'}])] \\ &= \text{cov}(E[X_{i,j} | \Theta_i], E(X_{i,j'} | \Theta_i)) \end{aligned}$$

Pour le troisième terme (5),

$$\begin{aligned} & E[(X_{i,j} - E[X_{i,j}|\Theta_i]) (E[X_{i,j'}|\Theta_i] - E[X_{i,j'}])] \\ &= E[E[(E[X_{i,j}|\Theta_i] - E[X_{i,j}]) (X_{i,j'} - E[X_{i,j'}|\Theta_i]) | \Theta_i]] \\ &= E[(E[X_{i,j}|\Theta_i] - E[X_{i,j}]) E[X_{i,j'} - E[X_{i,j'}|\Theta_i] | \Theta_i]] \end{aligned}$$

avec $E[X_{i,j'} - E[X_{i,j'}|\Theta_i] | \Theta_i] = 0$ donc

$$E[(X_{i,j} - E[X_{i,j}|\Theta_i]) (E[X_{i,j'}|\Theta_i] - E[X_{i,j'}])] = 0.$$

On démontre de même que le quatrième terme (6) est nul

$$E[(E[X_{i,j}|\Theta_i] - E[X_{i,j}]) (X_{i,j'} - E[X_{i,j'}|\Theta_i])] = 0.$$

Conséquences On se place sous les hypothèses (1-4) du modèle de Bühlmann.

1. Si $j = j'$. On obtient la **formule de la variance totale**

$$\begin{aligned} \text{var}(X_{i,j}) &= E[\text{var}(X_{i,j}|\Theta_i)] + \text{var}(E[X_{i,j}|\Theta_i]) \\ &= \Sigma^2 + M^2. \end{aligned}$$

Interprétation : la variance des sinistres se décompose comme la somme de la variance intra-contrat et de la variance inter-contrat.

2. Si $j \neq j'$. On a supposé que $\text{cov}(X_{i,j}, X_{i,j'} | \Theta_i) = 0$ et on a donc

$$\text{cov}(X_{i,j}, X_{i,j'}) = \text{cov}(E[X_{i,j}|\Theta_i], E[X_{i,j'}|\Theta_i]).$$

D'après les hypothèses du modèle de Bühlmann, on a $E[X_{i,j}|\Theta_i] = E[X_{i,j'}|\Theta_i]$. On en déduit alors que

$$\text{cov}(X_{i,j}, X_{i,j'}) = M^2.$$

Interprétation : la covariance entre les sinistres pour deux années distinctes d'un contrat est égale à la variance inter-contrat.

Exercice 3 . On reprend les hypothèses de l'exercice 1 et on note $\pi = P(\Theta_i = B) = \frac{3}{4}$, $p_1 = P(N_{i,j} = 1 | \Theta_i = B) = \frac{1}{15}$ et $p_2 = P(N_{i,j} = 1 | \Theta_i = M) = \frac{1}{10}$ avec $N_{i,j}$ la variable binaire qui décrit la survenance d'un sinistre. Les montants des sinistres sont donnés par $X_{i,j} = 1000N_{i,j}$. On suppose que les hypothèses du modèle de Bühlmann s'appliquent.

1. Calculer M^2 en fonction de π , p_1 et p_2 .
2. Calculer Σ^2 en fonction de π , p_1 et p_2 .
3. Calculer $\text{var}(X_{i,j})$ et vérifier que la formule de la variance totale s'applique.
4. Calculer $\text{cov}(X_{i,j}, X_{i,j'})$.
5. Ecrire une fonction `R` `sim=function(pi,p1,p2,m=20,n=10)` qui permet de simuler les sinistres pour un portefeuille de m contrats pendant n années. La fonction renverra une matrice X de dimension m par n (par défaut $m = 20$ et $n = 10$).
6. Vérifier les formules des questions 3. et 4. par simulation. On pourra tester différentes valeurs de p_1 et p_2 .

1. On a $E[N_{i,j}|\Theta_i = B] = p_1$ (espérance d'une loi de Bernoulli de paramètre p_1) et $E[N_{i,j}|\Theta_i = M] = p_2$. On en déduit donc que

$$E(N_{i,j}|\Theta_i) = p_1 \mathbf{1}(\Theta_i = B) + p_2 \mathbf{1}(\Theta_i = M).$$

$E(N_{i,j}|\Theta_i)$ est donc une variable aléatoire discrètes à valeurs dans $\{p_1, p_2\}$ dont il faut calculer la variance. Pour cela, on peut remarquer que

$$E(N_{i,j}|\Theta_i) = p_1 + (p_2 - p_1)Z$$

avec $Z = \mathbf{1}(\Theta_i = M) \sim \text{Ber}(1 - \pi)$. On a $\text{var}(Z) = \pi(1 - \pi)$ et on déduit que $\text{var}(E(N_{i,j}|\Theta_i)) = \pi(1 - \pi)(p_1 - p_2)^2$. Finalement $M^2 = \text{var}(E(X_{i,j}|\Theta_i)) = \text{var}(E(1000N_{i,j}|\Theta_i)) = 1000^2 \pi(1 - \pi)(p_1 - p_2)^2$.

2. $\text{var}(N_{i,j}|\Theta_i) = E[N_{i,j}^2|\Theta_i] - E[N_{i,j}|\Theta_i]^2$ avec

- $E(N_{i,j}^2|\Theta_i) = p_1 \mathbf{1}(\Theta_i = B) + p_2 \mathbf{1}(\Theta_i = M)$ (car $N_{i,j}^2 = N_{i,j}$)
- $E(N_{i,j}|\Theta_i)^2 = p_1^2 \mathbf{1}(\Theta_i = B) + p_2^2 \mathbf{1}(\Theta_i = M)$.

Donc $\text{var}(N_{i,j}|\Theta_i) = p_1(1 - p_1) \mathbf{1}(\Theta_i = B) + p_2(1 - p_2) \mathbf{1}(\Theta_i = M)$. On en déduit que $E(\text{var}(N_{i,j}|\Theta_i)) = \pi p_1(1 - p_1) + (1 - \pi)p_2(1 - p_2)$. Et enfin

$$\Sigma^2 = E(\text{var}(X_{i,j}|\Theta_i)) = E(\text{var}(1000X_{i,j}|\Theta_i)) = 1000^2 (\pi p_1(1 - p_1) + (1 - \pi)p_2(1 - p_2)).$$

3. $N_{i,j} \sim \text{Ber}(p)$ avec $p = P(N_{i,j} = 1) = \pi p_1 + (1 - \pi)p_2$. Donc $\text{var}(N_{i,j}) = p(1 - p)$ et $\text{var}(X_{i,j}) = 1000^2 p(1 - p)$. Pour montrer la formule de la variance totale, il suffit de vérifier que $\text{var}(N_{i,j}) = \text{var}(E(N_{i,j}|\Theta_i)) + E(\text{var}(N_{i,j}|\Theta_i))$. On a

$$\begin{aligned} \text{var}(N_{i,j}) &= (\pi p_1 + (1 - \pi)p_2)(\pi(1 - p_1) + (1 - \pi)(1 - p_2)) \\ &= \pi^2 p_1(1 - p_1) + (1 - \pi)^2 p_2(1 - p_2) + \pi(1 - \pi)(p_1(1 - p_2) + p_2(1 - p_1)) \\ &= \pi^2 p_1(1 - p_1) + (1 - \pi)^2 p_2(1 - p_2) + \pi(1 - \pi)((p_1 - p_2)^2 + p_1(1 - p_1) + p_2(1 - p_2)) \\ &= p_1(1 - p_1)(\pi^2 + \pi(1 - \pi)) + p_2(1 - p_2)((1 - \pi)^2 + \pi(1 - \pi)) + \pi(1 - \pi)(p_1 - p_2)^2 \\ &= p_1(1 - p_1)\pi + p_2(1 - p_2)(1 - \pi) + \pi(1 - \pi)(p_1 - p_2)^2 \\ &= \text{var}(E(N_{i,j}|\Theta_i)) + E(\text{var}(N_{i,j}|\Theta_i)) \end{aligned}$$

4. D'après le cours, on a $\text{cov}(X_{i,j}, X_{i,j'}) = M^2$ si $j \neq j'$ avec M^2 donné dans la question 1.

5. cf code R ci-dessous.

#Question 5

```
sim=function(pi,p1,p2,m=20,n=10){
  X=matrix(0,m,n)  #initialisation
  N=X
  theta=NULL
  for(i in 1:m){
    theta[i] = rbinom(1,size = 1,prob = pi)  #simulation profil de risque
    #theta=1 si Bon conducteur, theta=0 si Mauvais conducteur
    N[i,]=rbinom(n,size=1,p1*theta[i]+p2*(1-theta[i]))  #simulation des sinistres pour les n années
    #p1*theta[i]+p2*(1-theta[i]) =p1 si theta[i]=1
    #p1*theta[i]+p2*(1-theta[i]) =p2 si theta[i]=0
  }
  X=1000*N
}
```

```

    return(X)
}

#Question 6
#simulation du portefeuille
#p1=1/15 #vous pouvez changer les valeurs de p1 et p2
#p2=1/10
p1 = 1/2
p2 = 1/100
pi=3/4
X=sim(pi=pi,p1=p1,p2=p2,m=10^5,n=5)

M2=1000^2 *pi *(1-pi) *(p1-p2)^2 #valeurs théoriques
Sigma2=1000^2*(pi*p1*(1-p1)+(1-pi)*p2*(1-p2))

#Vérifions que var(X1)=M2+Sigma2
var(X) #matrice de variance covariance

```

```

##          [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 235123.23 44396.20 44459.15 45161.20 44761.39
## [2,] 44396.20 235132.99 46094.02 46196.10 45006.35
## [3,] 44459.15 46094.02 235298.37 46429.30 45550.73
## [4,] 45161.20 46196.10 46429.30 235032.83 44840.48
## [5,] 44761.39 45006.35 45550.73 44840.48 234608.98

```

```

M2+Sigma2 #valeur proche de la diagonale de var(X)

```

```

## [1] 234993.8

```

```

#Vérifions que cov(X1,X2)=M2
M2 ##valeur proche des éléments hors diagonale de var(X)?

```

```

## [1] 45018.75

```

```

#La valeur empirique de cov(X1,X2) fluctue quand on répète les simulations
# Utiliser des valeurs de p1 et p2 significativement différentes permet d'augmenter la corrélation

```

3.3 Calcul de la prime linéaire optimale

La proposition suivante donne le résultat principal du cours.

Proposition 2 *Sous les hypothèses du modèle de Bühlmann, les poids solutions du problème d'optimisation (2) sont donnés par*

- $c_0 = \frac{\Sigma^2}{\Sigma^2 + nM^2} \mu,$
- $c_k = \frac{M^2}{\Sigma^2 + nM^2}$ pour $k \geq 1$.

La prime de crédibilité linéaire (ou prime de Bühlmann) s'écrit donc

$$p_{i,n+1} = \frac{\Sigma^2}{\Sigma^2 + nM^2}\mu + \frac{nM^2}{\Sigma^2 + nM^2}\bar{X}_i \quad (7)$$

avec $\bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{i,j}$.

Preuve 2 Etudions les points critiques de la fonction

$$F(\lambda_0, \lambda_1, \dots, \lambda_n) = E[(X_{i,n+1} - (\lambda_0 + \lambda_1 X_{i,1} + \dots + \lambda_n X_{i,n}))^2]$$

Pour simplifier le calcul les dérivées partielles, on peut intervertir espérance et dérivée (on peut démontrer que c'est possible en utilisant les théorèmes d'interversion classiques). On a alors

$$\begin{aligned} \frac{\partial F}{\partial \lambda_0}(\lambda_0, \lambda_1, \dots, \lambda_n) &= \frac{\partial}{\partial \lambda_0} E[(X_{i,n+1} - (\lambda_0 + \lambda_1 X_{i,1} + \dots + \lambda_n X_{i,n}))^2] \\ &= E\left[\frac{\partial}{\partial \lambda_0} (X_{i,n+1} - (\lambda_0 + \lambda_1 X_{i,1} + \dots + \lambda_n X_{i,n}))^2\right] \\ &= -2E[(X_{i,n+1} - (\lambda_0 + \lambda_1 X_{i,1} + \dots + \lambda_n X_{i,n}))] \\ &= 2 \left(\mu \left(\sum_{j=1}^n \lambda_j - 1 \right) + \lambda_0 \right) \end{aligned}$$

car $E[X_{i,j}] = \mu$.

De même, pour $k \in \{1, \dots, n\}$, on obtient

$$\begin{aligned} \frac{\partial F}{\partial \lambda_k}(\lambda_0, \lambda_1, \dots, \lambda_n) &= E\left[\frac{\partial}{\partial \lambda_k} (X_{i,n+1} - (\lambda_0 + \lambda_1 X_{i,1} + \dots + \lambda_n X_{i,n}))^2\right] \\ &= -2E[X_{i,k} (X_{i,n+1} - (\lambda_0 + \lambda_1 X_{i,1} + \dots + \lambda_n X_{i,n}))] \\ &= -2(E[X_{i,k} X_{i,n+1}] - \lambda_0 E[X_{i,k}] - \lambda_1 E[X_{i,k} X_{i,1}] - \dots - \lambda_n E[X_{i,k} X_{i,n}]) \end{aligned}$$

Par hypothèse, on a

$$E[X_{i,j}, X_{i,j'}] = \begin{cases} M^2 + \Sigma^2 + \mu^2 & \text{si } j = j' \\ M^2 + \mu^2 & \text{si } j \neq j' \end{cases}$$

et donc en déduit que

$$\frac{\partial F}{\partial \lambda_k}(\lambda_0, \lambda_1, \dots, \lambda_n) = 2 \left(\lambda_0 \mu + (M^2 + \mu^2) \left(\sum_{j=1}^n \lambda_j - 1 \right) + \Sigma^2 \lambda_k \right).$$

Soit $(c_0, c_1, \dots, c_n)$ un point critique de F c'est à dire tel que $\frac{\partial F}{\partial \lambda_k}(c_0, c_1, \dots, c_n) = 0$ pour $k \in \{0, \dots, n\}$. $(c_0, c_1, \dots, c_n)$ est solution d'un système d'équations linéaires à $n+1$ inconnues.

En particulier, on a pour $k \in \{2, \dots, n\}$,

$$\frac{\partial F}{\partial \lambda_k}(c_0, c_1, \dots, c_n) - \frac{\partial F}{\partial \lambda_1}(c_0, c_1, \dots, c_n) = 2\Sigma^2(c_1 - c_k) = 0$$

et donc $c_1 = c_k$ (on suppose que $\Sigma^2 \neq 0$). Des égalités $\frac{\partial F}{\partial \lambda_0}(c_0, c_1, \dots, c_n) = 0$ et $\frac{\partial F}{\partial \lambda_1}(c_0, c_1, \dots, c_n) = 0$ on déduit finalement que

$$\begin{aligned} c_0 &= \frac{\Sigma^2}{\Sigma^2 + nM^2}\mu \\ c_k &= \frac{M^2}{\Sigma^2 + nM^2} \text{ si } k \geq 1. \end{aligned}$$

Réciproquement, on montre que la solution identifiée ci-dessus est bien un point critique de F . On peut vérifier que c'est un minimum global par un argument de convexité.

Interprétation. On a

$$p_{i,n+1} = \frac{nM^2}{\Sigma^2 + nM^2} \bar{X}_i + \frac{\Sigma^2}{\Sigma^2 + nM^2} \mu = \lambda(n, M^2, \Sigma^2) \bar{X}_i + (1 - \lambda(n, M^2, \Sigma^2)) \mu$$

avec $\lambda(n, M^2, \Sigma^2) = \frac{nM^2}{\Sigma^2 + nM^2} = \frac{1}{\frac{\Sigma^2}{nM^2} + 1}$ le poids (facteur de crédibilité) donné à la sinistralité historique du i ème assuré. On vérifie que ce facteur de crédibilité varie de manière conforme à l'intuition. En particulier $\lambda(n, M^2, \Sigma^2)$

- augmente avec n et tend vers 1 lorsque $n \rightarrow +\infty$: on donne de plus en plus de poids à la sinistralité passée lorsque la durée de l'historique augmente.
- diminue lorsque $\frac{\Sigma^2}{M^2}$ augmente, c'est à dire lorsque l'hétérogénéité du portefeuille diminue. En particulier si le portefeuille est homogène ($M^2 = \text{var}(E[X_{i,j}|\Theta_i]) = 0$) alors $\lambda(n, M^2, \Sigma^2) = 0$ et la prime n'est pas individualisée. A l'opposé, si la variance intra-contrat est nulle ($\Sigma^2 = E[(X_{i,j} - E[X_{i,j}|\Theta_i])^2] = 0$) alors $\lambda(n, M^2, \Sigma^2) = 1$.

3.4 Estimation des paramètres

En pratique, lorsqu'on veut utiliser la formule (7) pour calculer la prime de crédibilité linéaire, il faut pouvoir estimer les paramètres inconnus μ , Σ^2 et M^2 . Pour cela, on utilise généralement les estimateurs définis ci-dessous, qui sont sans biais sous les hypothèses (1-4) du modèle de Bühlmann :

- **Estimation de la prime collective μ .** $\bar{X} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n X_{i,j}$ est un estimateur sans biais de μ .
- **Estimation de la variance intra-contrat $\Sigma^2 = E[(X_{i,j} - E[X_{i,j}|\Theta_i])^2]$.** On note respectivement $\bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{i,j}$ et $S_i^2 = \frac{1}{n-1} \sum_{j=1}^n (X_{i,j} - \bar{X}_i)^2$ la moyenne et la variance empirique du contrat i . Un estimateur sans biais de Σ^2 est $\hat{\Sigma}^2 = \frac{1}{m} \sum_{i=1}^m S_i^2$ (moyenne des variances empiriques des différents portefeuilles).
- **Estimation de la variance inter-contrat $M^2 = \text{var}(E[X_{i,j}|\Theta_i])$.** Un estimateur sans biais de M^2 est

$$\hat{M}^2 = \frac{1}{m-1} \sum_{i=1}^m (\bar{X}_i - \bar{X})^2 - \frac{\hat{\Sigma}^2}{n}$$

($\frac{1}{m-1} \sum_{i=1}^m (\bar{X}_i - \bar{X})^2$ est la variance empirique des sinistres moyens sur les différents portefeuilles, $\frac{\hat{\Sigma}^2}{n}$ un terme correctif pour obtenir un estimateur sans biais). Cet estimateur prend parfois des valeurs négatives pour les petits échantillons...on peut alors utiliser l'estimateur biaisé $\tilde{M}^2 = \frac{1}{m-1} \sum_{i=1}^m (\bar{X}_i - \bar{X})^2$.

Remarque. Si on remplace μ , Σ^2 et M^2 par leurs estimateurs dans (7), on obtient la prime de crédibilité linéaire utilisée en pratique

$$\hat{p}_{i,n+1} = (1 - \hat{\lambda}) \bar{X} + \hat{\lambda} \bar{X}_i$$

avec $\hat{\lambda} = \frac{n\hat{M}^2}{\hat{\Sigma}^2 + n\hat{M}^2}$. Cette prime s'interprète bien comme une combinaison linéaire entre le prime collective empirique $\bar{X}$ et la prime individuelle collective $\bar{X}_i$. Le facteur de crédibilité $\hat{\lambda}$ dépend du nombre d'années n et de l'hétérogénéité du portefeuille mesurée par le rapport entre la variance intra-contrat et la variance inter-contrat.

Exercice 4 On reprend les hypothèses des exercices 1 et 3.

1. Ecrire une fonction R `estim=function(X)` qui calcule les estimateurs sans biais de μ , Σ^2 et M^2 définis dans le cours, avec X une matrice qui contient les sinistres.
2. Vérifier, en utilisant des simulations, que les estimateurs sont bien sans biais. On pourra tester différentes valeurs de p_1 et p_2 .

```
#question 1
#estimation des paramètres
estim=function(X){
  S2=apply(X,1,var) #variances par assuré
  barxi=apply(X,1,mean) #moyenne des sinistres pour chaque assuré
  Sig2=mean(S2) #Estimateur de Sigma^2
  M2=var(barxi)-Sig2/ncol(X) #Estimateur de M^2
  return(list(mu=mean(X),Sig2=Sig2,M2=M2))
}

#test
#simulation du portefeuille
p1=1/100 #vous pouvez changer les valeurs de p1 et p2
p2=1/2
pi=3/4
X=sim(pi=pi,p1=p1,p2=p2)
estim(X)
```

```
## $mu
## [1] 140
##
## $Sig2
## [1] 63333.33
##
## $M2
## [1] 60403.51
```

```
#question 2
#étude des estimateurs de M2 par simulation

#valeurs théoriques
M2vrai=1000^2 *pi *(1-pi) *(p1-p2)^2
Sig2vrai=1000^2*(pi*p1*(1-p1)+(1-pi)*p2*(1-p2))
muvrai=1000*(p1*pi+p2*(1-pi))

#simulation de la loi des estimateurs
N=10^4

M2sim=NULL
Sig2sim=NULL
musim=NULL
for (i in 1:N){
  res=estim(sim(pi=pi,p1=p1,p2=p2))
```

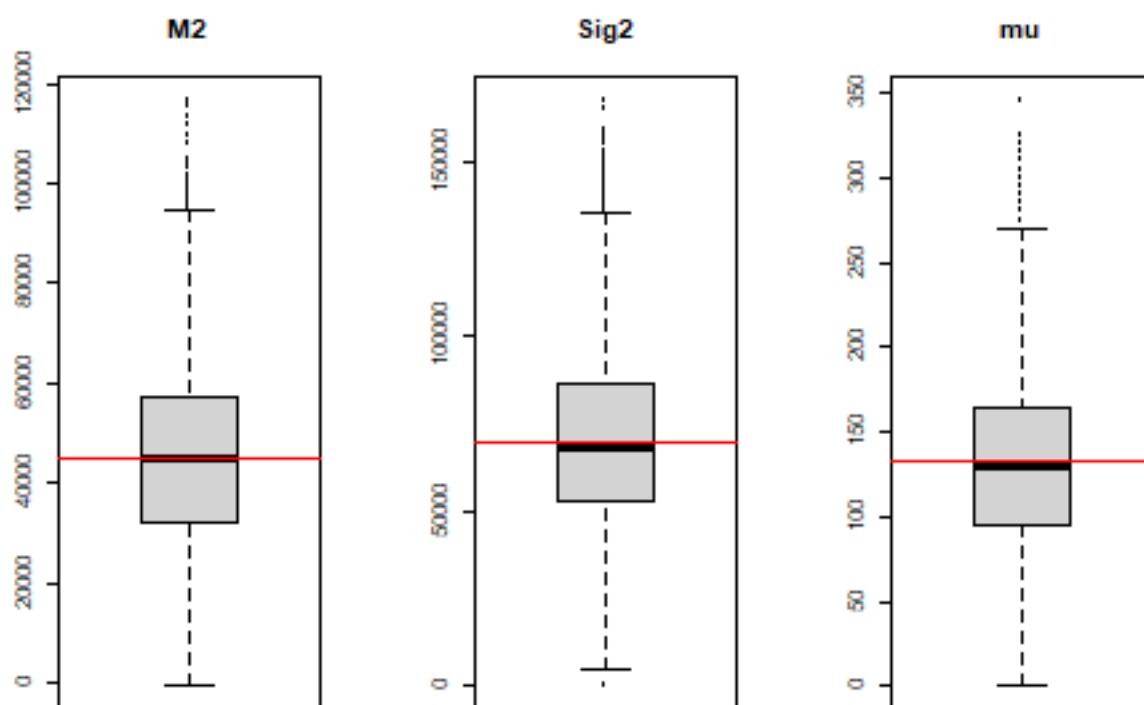
```

M2sim[i]=res$M2
Sig2sim[i]=res$Sig2
musim[i]=res$mu
}

par(mfrow=c(1,3))

boxplot(M2sim,main='M2') #loi de l'estimateur
abline(h=M2vrai,col='red') #vrai valeur
boxplot(Sig2sim,main='Sig2') #loi de l'estimateur
abline(h=Sig2vrai,col='red') #vrai valeur
boxplot(musim,main='mu') #loi de l'estimateur
abline(h=muvrai,col='red') #vrai valeur

```



```

#estimation du biais de l'estimateur (doit être proche de 0 si sans biais)
mean(M2sim)-M2vrai

```

```
## [1] -78.75468
```

```
mean(Sig2sim)-Sig2vrai
```

```
## [1] -22.61111
```

```
mean(musim)-muvrai
```

```
## [1] -0.1345
```

```
#on peut calculer des IC à 95% pour voir si la différence est significative  
c(mean(M2sim)-1.96*sd(M2sim)/sqrt(N),mean(M2sim)+1.96*sd(M2sim)/sqrt(N)) #IC pour M2
```

```
## [1] 44580.73 45299.26
```

```
M2vrai #vraie valeur
```

```
## [1] 45018.75
```

Exercice 5 Le tableau 1 décrit le montant des sinistres pour 10 contrats pendant 10 ans.

1. Vérifier à l'aide d'une analyse de la variance à un facteur si on peut supposer que le portefeuille est homogène.
2. Donner une estimation de μ , Σ et M .
3. En déduire une estimation du facteur de crédibilité et de la prime de crédibilité linéaire pour l'année 11 pour les 10 assurés du tableau 1.

```
y=c(  
  0,0,10,0,0,0,0,5,0,0,  
  15,0,0,12,0,0,5,10,0,0,  
  0,0,2,0,0,0,3,0,5,0,  
  0,20,0,0,15,0,0,25,0,0,  
  0,0,0,0,5,0,0,10,0,0,  
  0,0,0,3,0,0,2,0,0,5,  
  5,0,0,0,2,0,5,0,0,0,  
  0,30,0,15,0,0,20,0,0,20,  
  0,0,2,0,0,0,5,0,2,0,  
  0,0,0,0,0,0,0,0,10,0)  
  
m=10  
x=matrix(y,nrow=m,byrow=TRUE); #matrice avec les sinistres  
  
z=matrix(as.factor(rep(1:m,10)),nrow=m) #matrice avec les numéros de contrat  
X=as.vector(x) #vecteur avec les sinistres  
Z=as.vector(z) #vecteur avec les numéros de contrat  
  
fit=lm(X~Z) #ajustement d'un modèle d'analyse de la variance à un facteur  
summary(fit) #p-value=0.025 < alpha=5% : le portefeuille est hétérogène  
  
##  
## Call:  
## lm(formula = X ~ Z)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##    -8.5    -1.5    -1.0     1.0    21.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.500e+00  1.772e+00   0.847  0.39949
## Z10          -5.000e-01  2.506e+00  -0.200  0.84229
## Z2           2.700e+00  2.506e+00   1.077  0.28414
## Z3          -5.000e-01  2.506e+00  -0.200  0.84229
## Z4           4.500e+00  2.506e+00   1.796  0.07588 .
## Z5           1.483e-14  2.506e+00   0.000  1.00000
## Z6          -5.000e-01  2.506e+00  -0.200  0.84229
## Z7          -3.000e-01  2.506e+00  -0.120  0.90497
## Z8           7.000e+00  2.506e+00   2.794  0.00637 **
## Z9          -6.000e-01  2.506e+00  -0.239  0.81131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.603 on 90 degrees of freedom
## Multiple R-squared:  0.1842, Adjusted R-squared:  0.1027
## F-statistic: 2.258 on 9 and 90 DF,  p-value: 0.02502
```

```
#alternatives
#la commande stack permet de créer directement les vecteurs X et Y
fit=lm(values~ind,data=stack(as.data.frame(t(x))))
#la fonction aov donne les mêmes résultats que lm (en plus interprétable?)
summary(aov(values~ind,data=stack(as.data.frame(t(x)))))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## ind           9  638.2   70.91   2.258  0.025 *
## Residuals    90 2825.6   31.40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#####
#Question 2
#on peut utiliser la fonction est de l'exercice 3.2
par=estim(x)
par
```

```
## $mu
## [1] 2.68
##
## $Sig2
## [1] 31.39556
##
## $M2
## [1] 3.951111
```

```
#####
#Question 3

n=ncol(x)
lambda2=n*par$M2/(par$Sig2+n*par$M2) #facteur de crédibilité (0.557)
barxi=apply(x,1,mean) #moyenne des sinistres pour chaque assuré
p=barxi*lambda2+par$mu*(1-lambda2) #Prime de crédibilité linéaire
cbind(x,prime_ind=barxi,prime_col=par$mu,prime_lin=p) #affichage des résultats
```

```
##
## prime_ind prime_col prime_lin
## [1,] 0 0 10 0 0 0 0 5 0 0 1.5 2.68 2.022472
## [2,] 15 0 0 12 0 0 5 10 0 0 4.2 2.68 3.526985
## [3,] 0 0 2 0 0 0 3 0 5 0 1.0 2.68 1.743859
## [4,] 0 20 0 0 15 0 0 25 0 0 6.0 2.68 4.529994
## [5,] 0 0 0 0 5 0 0 10 0 0 1.5 2.68 2.022472
## [6,] 0 0 0 3 0 0 2 0 0 5 1.0 2.68 1.743859
## [7,] 5 0 0 0 2 0 5 0 0 0 1.2 2.68 1.855304
## [8,] 0 30 0 15 0 0 20 0 0 20 8.5 2.68 5.923061
## [9,] 0 0 2 0 0 0 5 0 2 0 0.9 2.68 1.688136
## [10,] 0 0 0 0 0 0 0 0 10 0 1.0 2.68 1.743859
```

```
#La 11eme colonne donne la prime individuelle
#La 12eme colonne donne la prime collective
#la 13eme colonne donne la prime de crédibilité linéaire
#la prime de crédibilité linéaire est une moyenne pondérée
#entre la prime individuelle et la prime collective
#ici lambda2 est légèrement supérieur à 0.5 : on donne un peu plus de poids
#à la prime individuelle que à la prime collective
```

Exercice 6 (examen 2021-2022) Un portefeuille d'assurance compte deux classes d'assurés, notées A et B. 60% des contrats appartiennent à la classe A et 40% à la classe B. On suppose que la distribution du montant total des sinistres annuels suit une loi de Gauss dans chacune des classes. La moyenne et l'écart-type de ces lois de Gauss sont donnés dans le tableau suivant.

Classe	Proportion	Moyenne des sinistres	Ecart-type des sinistres
A	60%	100	100
B	40%	300	200

Table 3: Caractéristiques du portefeuille

On suppose que les hypothèses de Bühlmann s'appliquent et on reprend les notations du cours.

1. Calculer les quantités M^2 , Σ^2 et μ . Détailler les calculs sur la copie.
2. Calculer la prime pure d'un contrat pour la troisième année sachant qu'il a subi des sinistres de 200 puis 300 sur les deux premières années. Détailler les calculs sur la copie.
3. Calculer la prime de crédibilité linéaire d'un contrat pour la troisième année sachant qu'il a subi des sinistres de 200 puis 300 sur les deux premières années. Détailler les calculs sur la copie.

(a) Est-ce qu'on retrouve le même résultat que dans la question précédente? Commenter.

(b) Quel est le facteur de crédibilité λ ? Comment s'interprète cette quantité?

4. Ecrire un code R qui permet de simuler un portefeuille ayant les caractéristiques du tableau 3 puis vérifier les résultats de la question 1. en utilisant des simulations.

1. On note $\mu_1 = 100$, $V_1 = 100^2$, $\mu_2 = 300$ et $V_2 = 200^2$ les moyennes et variances des lois gaussiennes dans les groupes A et B. On note $\pi_1 = 0.6$ et $\pi_2 = 0.4$ les probabilités d'appartenance aux groupes A et B. On note $\theta_i \in \{1, 2\}$ le groupe de l'assuré ($\theta_i = 1$ si l'assuré i est dans le groupe A et $\theta_i = 2$ si l'assuré i est dans le groupe B). En reprenant les notations du cours, d'après les hypothèses de l'énoncé, on a $E[X_{i,j}|\theta_i = 1] = \mu_1$ et $E[X_{i,j}|\theta_i = 2] = \mu_2$, et donc, d'après la définition de l'espérance conditionnelle

$$E[X_{i,j}|\theta_i] = \mu_1 \mathbf{1}(\theta_i = 1) + \mu_2 \mathbf{1}(\theta_i = 2).$$

En utilisant les formules usuelles sur la moyenne et la variance des lois discrètes, on en déduit que

$$\mu = E[X_{i,j}] = \pi_1 \mu_1 + \pi_2 \mu_2$$

$$M^2 = \text{var}(E[X_{i,j}|\theta_i]) = \pi_1 (\mu_1 - \mu)^2 + \pi_2 (\mu_2 - \mu)^2.$$

On a

$$\text{var}(X_{i,j}|\theta_i) = V_1 \mathbf{1}(\theta_i = 1) + V_2 \mathbf{1}(\theta_i = 2)$$

et donc

$$\Sigma^2 = E[\text{var}(X_{i,j}|\theta_i)] = \pi_1 V_1 + \pi_2 V_2.$$

Les applications numériques sont données dans les codes ci-dessous.

2. Par définition, la prime pure est donnée par

$$E[X_{i,3}|X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}]$$

avec $x_{i,1} = 200$ et $x_{i,2} = 300$. On a

$$\begin{aligned} E[X_{i,3}|X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}] &= E[X_{i,3}|X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}, \theta_i = 1]P(\theta_i = 1|X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}) \\ &\quad + E[X_{i,3}|X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}, \theta_i = 2]P(\theta_i = 2|X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}) \end{aligned}$$

puis, en supposant que les sinistres sont conditionnellement indépendants connaissant les profils de risque (cf chapitre 2 du cours),

$$\begin{aligned} E[X_{i,3}|X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}] &= E[X_{i,3}|\theta_i = 1]P(\theta_i = 1|X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}) \\ &\quad + E[X_{i,3}|\theta_i = 2]P(\theta_i = 2|X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}) \end{aligned}$$

avec pour $k \in \{1, 2\}$, $E[X_{i,3}|\theta_i = k] = \mu_k$. De plus, d'après la formule de Bayes,

$$P(\theta_i = k|X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}) = \frac{P(X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}|\theta_i = k)P(\theta_i = k)}{P(X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2})}$$

et

$$\begin{aligned} P(X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}|\theta_i = k) &= P(X_{i,1} = x_{i,1}|\theta_i = k)P(X_{i,2} = x_{i,2}|\theta_i = k) \\ &= \phi(x_{i,1}; \mu_k, V_k)\phi(x_{i,2}; \mu_k, V_k) \end{aligned}$$

avec $\phi(\cdot; \mu, V)$ la densité de la loi normale de moyenne μ et variance V . De plus, on a

$$\begin{aligned} P(X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}) &= P(X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}|\theta_i = 1)P(\theta_i = 1) \\ &\quad + P(X_{i,1} = x_{i,1}, X_{i,2} = x_{i,2}|\theta_i = 2)P(\theta_i = 2). \end{aligned}$$

En utilisant les formules ci-dessus, on peut calculer la prime pure. Les applications numériques sont données dans les codes ci-dessous.

3. On utilise les formules du cours

$$p_{i,n+1} = \frac{\Sigma^2}{\Sigma^2 + nM^2} \mu + \frac{nM^2}{\Sigma^2 + nM^2} \bar{x}_i$$

avec μ , Σ^2 et M^2 calculés dans la question 1, $n = 2$, $\bar{x}_i = 250$. Les applications numériques sont données dans les codes ci-dessous.

On remarque que les primes des questions 2 et 3 sont différentes. La prime de la question 3 est optimale (au sens des moindres carrés) si on se restreint aux primes linéaires alors que la prime de la question 2 est optimale (au sens des moindres carrés) parmi toutes les primes possibles (linéaires et non-linéaires).

Le facteur de crédibilité est $\lambda = 0.47$. Il est proche de $1/2$: on donne un poids similaire à la prime collective et à la prime individuelle

4. Cf code ci-dessous.

```
pi=c(.6,.4) #proba d'appartenance aux groupes notées pi_1 et pi_2
mui=c(100,300) #moyennes des gaussiennes notées mu_1 et mu_2
Vi=c(100^2,200^2) #variances des gaussiennes notées V_1 et V_2
#question 1
mu=sum(pi*mui) # mu = pi_1 \mu_1 + pi_2 \mu_2
M2=sum(pi*(mui-mu)^2) # M^2 = pi_1 (\mu_1-\mu)^2 + pi_2 (\mu_2-\mu)^2
#on a aussi M^2 = pi_1*pi_2 (\mu_1-\mu_2)^2
Sigma2=sum(pi*Vi) # Sig^2 = pi_1 V_1 + pi_2 V_2
mu
```

```
## [1] 180
```

```
M2
```

```
## [1] 9600
```

```
Sigma2
```

```
## [1] 22000
```

```
#Question 2
x=c(200,300)
prob=rep(0,2)
for (k in 1:2){
  prob[k]=pi[k]*prod(dnorm(x,mean=mui[k],sd=sqrt(Vi[k])))
  #calcul de P(X_{i,1}=200,X_{i,2}=300,theta_i=k)
}

probn=prob/sum(prob)
#P(theta_i=k|X_{i,1}=200,X_{i,2}=300)
#=P(theta_i=k,X_{i,1}=1200,X_{i,2}=300)/
#(P(theta_i=1,X_{i,1}=200,X_{i,2}=300)+P(theta_i=2,X_{i,1}=200,X_{i,2}=300))
prim=sum(probn*mui)
#E[X|X_{i,1}=200,X_{i,2}=300]
#=E[X|theta_i=1]P(theta_i=1|X_{i,1}=200,X_{i,2}=300) + #E[X|theta_i=2]P(theta_i=2|X_{i,1}=200,X_{i,2}=300)
prim #prime pure demandée
```

```
## [1] 228.3625
```

```
#question 3
```

```
n=2
```

```
lambda= n*M2/(n*M2+Sigma2) #calcul facteur crédibilité
```

```
lambda
```

```
## [1] 0.4660194
```

```
#proche de 1/2, on donne un poids similaire à la prime collective et à la prime individuelle  
lambda*500/n+(1-lambda)*mu #calcul prime année 3
```

```
## [1] 212.6214
```

```
#question 4
```

```
sim=function(pi,mui,Vi,m=10,n=10){  
  theta=sample(x=1:2,size=m,replace=TRUE,prob=pi)  
  m=length(theta)  
  X=matrix(0,m,n)  
  for (j in 1:n){  
    X[,j]=rnorm(n=m,mean=mui[theta],sd=sqrt(Vi[theta]))  
  }  
  return(X)  
}
```

```
X=sim(pi,mui,Vi,m=10^4,n=100)
```

```
#simulation d'un grand portefeuille pour obtenir des estimateurs avec faible variance
```

```
#estimateur sans biais de \mu, \Sigma, M2
```

```
est=function(X){  
  S2=apply(X,1,var) #variances par assurés  
  barxi=apply(X,1,mean) #moyenne des sinistres pour chaque assuré  
  Sig2=mean(S2) #Estimateur de Sigma^2  
  M2=var(barxi)-Sig2/ncol(X) #Estimateur de M^2  
  return(list(mu=mean(X),Sig2=Sig2,M2=M2))  
}
```

```
estim(X) #on doit retrouver des valeurs proches des valeurs calculées dans la question 1
```

```
## $mu
```

```
## [1] 179.8382
```

```
##
```

```
## $Sig2
```

```
## [1] 21979.21
```

```
##
```

```
## $M2
```

```
## [1] 9563.721
```

```
c(mu,Sigma2,M2)
```

```
## [1] 180 22000 9600
```