TP4 Sujet

February 17, 2023

1 Réduction de variance et calcul de sensibilités

Sur on problème de modélisation classique en assurance, on illustre l'importance de l'erreur relative puis deux techniques de réduction de variance:

- méthode par préconditionnement,
- échantillonage d'importance (important pour les événements rares).

Dans une deuxième partie on s'intéresse aux sensibilités et comment implémenter efficacement la méthode des différences finies avec la méthode de Monte Carlo.

[ ]:
```
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_theme()
from numpy.random import default_rng
rng = default_rng()
```

1.1 Charge sinistre et loi Poisson-composée

On définit la *charge sinistre totale* (sur une période T) par la variable aléatoire positive

$$S = \sum_{i=1}^{N} X_i$$

où N est une variable aléatoire à valeurs dans $\mathbf{N}$ représentant le nombre de sinistres sur la période T, et pour $i \geq 1$, X_i est une variable aléatoire à valeurs dans $\mathbf{R}_+$ représentant le coût du i-ème sinistre, avec la convention selon laquelle la somme est nulle si $N = 0$. Les $(X_i)_{i\geq 1}$ sont supposées indépendantes et identiquement distribuées, et indépendantes de N (indépendance fréquences - coûts).

Une modélisation classique est de considérer

- N de loi de Poisson de paramètre $\lambda > 0$,
- X_1 de loi log-normale de paramètres $\mu > 0$, $\sigma^2 > 0$, c'est à dire $X_1 = \exp(G_1)$ avec $G_1 \sim \mathcal{N}(\mu, \sigma^2)$.

Le but est d'estimer la **probabilité de dépassement** c'est à dire calculer la probabilité que la charge sinistre totale dépasse un seuil K:

$$p = \mathbf{P}[S > K] \quad \text{pour } K \text{ grand}$$

Dans la suite on prend $\lambda = 10$, $\mu = 0.1$ et $\sigma = 0.3$ et on considère plusieurs valeurs du seuil K.

1.1.1 Question: simulation de la charge sinistre totale

Ecrire une fonction `simu_S(size, mu, sigma, lambd)` qui renvoie un échantillon de taille `size` de réalisations indépendantes de S.

[ ]:

1.1.2 Question: représentation graphique

Représenter l'histogramme d'un échantillon de 100 000 réalisations de S et du seuil $K = 20$ par une ligne verticale rouge.

[ ]:

1.2 Estimateur Monte Carlo et erreur relative

Soit $p_n = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{S^{(j)}>K}$ l'estimateur Monte Carlo de $p = \mathbf{P}[S > K]$ où $(S^{(j)})_{j=1,\dots,n}$ est une suite *i.i.d.* de même loi que S.

On rappelle que:

- l'**erreur absolue** de l'estimateur Monte Carlo p_n est définie par $|p_n - p|$ et qu'avec probabilité 0.95 cette erreur est bornée par $e_n = 1.96\frac{\sigma_n}{\sqrt{n}}$ avec $\sigma_n^2 = p_n - p_n^2$,
- l'**erreur relative** de l'estimateur Monte Carlo est définie par $\frac{|p_n - p|}{p}$ que l'on majore avec probabilité 0.95 par $\frac{e_n}{p_n}$.

1.2.1 Question: erreur relative

Ecrire une fonction `relative_error` qui à partir d'un échantillon de S (de taille n) et d'une valeur de seuil K renvoie la probabilité p_n et l'erreur relative (plus exactement la borne $\frac{e_n}{p_n}$ à 95%).

Tracer l'erreur relative d'un échantillon de taille 100 000 en fonction de K pour K allant de 20 à 30. Comment interpréter cette courbe?

[ ]:

1.2.2 Question: Monte Carlo à précision fixée

Mettre en oeuvre un estimateur de Monte Carlo qui s'arrête dès que l'erreur relative est de 5%. On pourra par exemple introduire la variable aléatoire

$$\tau^{(m)} = \inf\{n \geq 1, e_{nm} \leq 0.05 p_{nm}\},$$

qui dépend d'un paramètre m fixé, par exemple $m = 10\,000$, et renvoyer $p_{\tau^{(m)}}$ ainsi que l'erreur relative et la taille de l'estimateur associé. Le paramètre m permet de recalculer l'estimateur et l'erreur uniquement toutes les m itérations et donc de réduire la complexité par rapport au choix naïf $m = 1$. On appelle ce paramètre m la taille du *batch* (size batch). Le nombre d'itérations (la taille de l'échantillon) dans la méthode de Monte Carlo pour un $\tau^{(m)}$ donné est donc $\tau^{(m)} \times m$.

Définir la fonction qui code cet estimateur Monte Carlo:

`monte_carlo_relative(mu, sigma, lambd, K, size_batch = 10000, error = 0.05)`

[ ]:

1.2.3 Question: complexité en fonction de K

Reproduire un tableau de résultat similaire au tableau suivant obtenu avec cet estimateur de Monte Carlo adaptatif jusqu'à l'itération $\tau^{(m)} \times m$ pour une erreur relative de 10% et pour différentes valeurs de $K = 20, \dots, 25$.

Tracer le nombre d'itérations nécessaires en fonction de K.

[1]:
```
import pandas as pd
df = pd.read_pickle("data/iterations_df.pkl")
df
```

[1]:

	Probabilité p_n	Erreur relative	Itérations
20	0.021600	0.093277	20000
21	0.013100	0.098219	30000
22	0.007820	0.098733	50000
23	0.003910	0.098927	100000
24	0.002259	0.099908	170000
25	0.001258	0.099186	310000

[ ]:

1.3 Réduction de variance par préconditionnement

Pour réduire la variance on teste d'abord l'idée présentée dans l'exercice 1 du TD3, c'est à dire qu'on considère la variable aléatoire

$$M = \inf\{r \geq 1, \sum_{i=1}^{r} X_i > K\}$$

et la représentation suivante

$$p = \mathbf{E}[\phi(M)] \quad \text{avec} \quad \phi(m) = \mathbf{P}[N \geq m]$$

1.3.1 Question: simulation de M

Ecrire une fonction `simu_M` similaire à la fonction `simu_S` avec l'argument K supplémentaire qui renvoie un échantillon *i.i.d.* de même loi que M.

[ ]:

1.3.2 Question: Monte Carlo et ratio de variance

En utilisant la fonction `monte_carlo` du TP précédent. Calculer le ratio de variance entre l'estimateur p_n et l'estimateur basé sur la représentation $p = \mathbf{E}[\phi(M)]$ où ϕ est calculée en utilisant la fonction de survie et la fonction de masse de la loi de Poisson (cf. la documentation de `stats.poisson`). Faire ce calcul pour différentes valeurs de K et $n = 20\,000$

[ ]:

1.4 Réduction de variance par échantillonage d'importance

Pour réduire la variance sans faire exploser la complexité pour les grandes valeurs de K on propose une méthode d'échantillonage d'importance (Importance Sampling) en modifiant la loi de la variable aléatoire N (on peut faire un autre choix, en changeant la loi des X_i ou bien en changeant la loi de N et des X_i). Le changement de loi proposé ici repose sur le changement de probabilité, pour $\theta \in \mathbf{R}$

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{P}_\theta} = L_\theta \quad \text{avec} \quad L_\theta = \exp(-\theta N + \psi(\theta)),$$

où $\psi(\theta) = \log \mathbf{E}\big[\exp(\theta N)\big] = \lambda(e^\theta - 1)$. On vérifie par le calcul que la loi de N sous $\mathbf{P}_\theta$ est la loi de Poisson de paramètre $\tilde{\lambda} = \lambda e^\theta$. Ainsi on a la représentation

$$\mathbf{P}\left[\sum_{i=1}^{N} X_i > K\right] = \mathbf{E}_{\mathbf{P}_\theta}\left[\mathbf{1}_{\sum_{i=1}^{N} X_i > K} \exp\big(-\theta N + \psi(\theta)\big)\right] \quad \text{avec } N \sim \mathcal{P}(\tilde{\lambda}) \text{ sous } \mathbf{P}_\theta.$$

Il est d'usage pour la loi de Poisson d'écrire la variable L_θ à partir de λ et $\tilde{\lambda}$ (la valeur du paramètre de la loi de Poisson sous la nouvelle probabilité) *i.e.*

$$L_\theta = \exp\big(-\theta N + \lambda(e^\theta - 1)\big) = \left(\frac{\lambda}{\tilde{\lambda}}\right)^N \exp\big(\tilde{\lambda} - \lambda\big).$$

1.4.1 Question: simulation sous $\mathbf{P}_\theta$

La loi de N sous $\mathbf{P}_\theta$ est la loi de Poisson de paramètre $\tilde{\lambda} = \lambda e^\theta$ et la suite $(X_i)_{i\geq 1}$ est indépendante de N donc de L_θ et n'est donc pas impactée par le changement de probabilité: la loi des $(X_i)_{i\geq 1}$ est inchangée.
Ecrire une fonction `simu_S_tilde` inspirée de `simu_S` qui prend un paramètre supplémentaire θ et qui renvoie un échantillon de $\sum_{i=1}^{N} X_i$ sous $\mathbf{P}_\theta$.

[ ]:

1.4.2 Question: Monte Carlo sous $\mathbf{P}_\theta$

Comparer pour différentes valeurs de K, avec $\theta = 0.7$, l'estimateur de Monte Carlo basé sur la représentation

$$\mathbf{P}\left[\sum_{i=1}^{N} X_i > K\right] = \mathbf{E}\left[\mathbf{1}_{\sum_{i=1}^{\tilde{N}} X_i > K} \left(\frac{\lambda}{\tilde{\lambda}}\right)^{\tilde{N}} \exp(\tilde{\lambda} - \lambda)\right] \quad \text{avec } \tilde{N} \sim \mathcal{P}(\tilde{\lambda}).$$

Pour $K = 22$ le ratio de variance est de l'ordre de 16-17.

Que se passe-t-il si le paramètre θ est mal choisi? (prendre par exemple $\theta = 1.2$ puis $\theta = 1.5$, et $\theta = -0.1$…)

[ ]:

1.5 Calcul de sensibilités

On utilisera la notation $S^{(\lambda)}$ pour indiquer la dépendance de variable aléatoire $S = \sum_{i=1}^{N} X_i$ en le paramètre $\lambda > 0$ (paramètre de la loi de Poisson sous-jacente). On s'intéresse à la sensibilité de la probabilité p en fonction de lambda c'est à dire

$$\frac{\partial}{\partial\lambda} p(\lambda) = \frac{\partial}{\partial\lambda} \mathbf{P}[S^{\lambda} > K]$$

1.5.1 Différences finies

Implémenter l'estimateur Monte Carlo basé sur les différences finies d'ordre 2

$$\frac{\partial}{\partial\lambda} p(\lambda) = \frac{p(\lambda + h) - p(\lambda - h)}{2h} + \mathcal{O}(h^2)$$

Comme vu en cours, il y a plusieurs façon d'implémenter l'estimateur Monte Carlo dans ce cadre biaisé.

- Le premier estimateur naïf $J_{n,h}^{(1)}(\lambda)$ est basé sur des réalisations indépendantes de $S^{(\lambda+h)}$ et $S^{(\lambda-h)}$ et n'est pas efficace: la variance explose lorsque h tend vers 0. Ainsi on pose

  $$J_{n,h}^{(1)}(\lambda) = \frac{1}{2hn}\Big(\sum_{k=1}^{n} \mathbf{1}_{\{S_k^{(\lambda+h)} > K\}} - \sum_{k=1}^{n} \mathbf{1}_{\{\tilde{S}_k^{(\lambda-h)} > K\}}\Big),$$

  où $(S_k^{(\lambda+h)})_{k\geq 1}$ et $(\tilde{S}_k^{(\lambda-h)})_{k\geq 1}$ sont des suites indépendantes de variables aléatoires *i.i.d.*.

- Le deuxième estimateur $J_{n,h}^{(2)}(\lambda)$ utilise des réalisations fortements corrélées de la loi de Poisson au sens suivant: on utilise la même réalisation uniforme U pour constuire deux réalisations $N^{(\lambda+h)}$ et $N^{(\lambda-h)}$ en utilisant la méthode de l'inverse de la fonction de répartition. Dans ce deuxième estimateur, les lois log-normales sont indépendantes. On a donc

  $$J_{n,h}^{(2)}(\lambda) = \frac{1}{2hn} \sum_{k=1}^{n} \Big(\mathbf{1}_{\{S_k^{(\lambda+h)} > K\}} - \mathbf{1}_{\{\bar{S}_k^{(\lambda-h)} > K\}}\Big),$$

  où pour $k \geq 1$, $S_k^{(\lambda+h)} = \sum_{i=1}^{G(\lambda+h,U_k)} X_{i,k}$ et $\bar{S}_k^{(\lambda-h)} = \sum_{i=1}^{G(\lambda-h,U_k)} \bar{X}_{i,k}$ avec $G(\lambda, u)$ l'inverse généralisée de la loi de Poisson de paramètre λ, $(U_k)_{k\geq 1}$ suite *i.i.d.* uniforme sur $[0, 1]$ indépendante de $(X_{i,k})_{i\geq 1,k\geq 1}$ et $(\bar{X}_{i,k})_{i\geq 1,k\geq 1}$ deux suites (doublement indicées) *i.i.d.* de loi log-normale (de paramètres μ et σ inchangés).

- Un troisième estimateur $J_{n,h}^{(3)}(\lambda)$ utilise des réalisations fortements corrélées de la loi de Poisson et des variables aléatoires log-normales communes.

  $$J_{n,h}^{(3)}(\lambda) = \frac{1}{2hn} \sum_{k=1}^{n} \Big(\mathbf{1}_{\{S_k^{(\lambda+h)} > K\}} - \mathbf{1}_{\{S_k^{(\lambda-h)} > K\}}\Big),$$

  où pour $k \geq 1$, $S_k^{(\lambda+h)} = \sum_{i=1}^{G(\lambda+h,U_k)} X_{i,k}$ et $S_k^{(\lambda-h)} = \sum_{i=1}^{G(\lambda-h,U_k)} X_{i,k}$ avec $G(\lambda, u)$ l'inverse généralisée de la loi de Poisson de paramètre λ, $(U_k)_{k\geq 1}$ suite *i.i.d.* uniforme sur $[0, 1]$ indépendante de $(X_{i,k})_{i\geq 1,k\geq 1}$ une suite (doublement indicée) *i.i.d.* de loi log-normale.

1.5.2 Question: plusieurs estimateurs des différences finies

On fixe les paramètres $\lambda = 10$, $\mu = 0.1$, $\sigma = 0.3$ et $K = 20$. Programmer ces 3 estimateurs pour différentes valeurs de h (par exemple, $h = 1$, 0.5, 0.1 et 0.01), et donner le résultat des estimateurs Monte Carlo avec $n = 50\,000$.

Que se passe-t-il lorsque h tend vers 0? Comparez le comportement pour ces 3 estimateurs. Il est très important de bien interpréter ces tableaux de résultats et de conclure qu'il faut utiliser l'estimateur $J^{(3)}_{n,h}(\lambda)$ et en aucun cas l'estimateur $J^{(1)}_{n,h}(\lambda)$.

Remarque: on considère ici uniquement l'étude de l'erreur statistique dûe à la méthode de Monte Carlo. On ne considère pas l'erreur de biais qui décroît lorsque h tend vers 0 et qui est peut-être non négligeable pour $h = 1$. Les IC construits ici sont biaisés et on ne peut pas affirmer que la vraie valeur est dans l'IC à 95% (au moins pour les grandes valeurs de h).

[ ]: