


Probabilités IV

STEP, MINES ParisTech*

29 janvier 2020 (#f301eed)

Table des matières

Lois conditionnelles dans un couple	3
Cas où X est discrète	3
Remarque	4
Exemple	5
Densités conditionnelles	6
Proposition	6
Proposition	7
Cas général	7
Théorème	7
Remarques	8
Conséquences	8
Proposition	8
Exemple	9
Proposition (critère d'indépendance)	10
Espérance conditionnelle	10
Définition	10
Remarques	11
Définition	11
Théorème	11
Proposition — transfert conditionnel	12
Vecteurs Gaussiens à densité	12
Régression et espérance conditionnelle des variables de carré intégrable	14

*Ce document est un des produits du projet  **boisgera**/CDIS, initié par la collaboration de (S)ébastien Boiségerault (CAOR), (T)homas Romary et (E)milie Chautru (GEOSCIENCES), (P)auline Bernard (CAS), avec la contribution de Gabriel Stoltz (Ecole des Ponts ParisTech, CERMICS). Il est mis à disposition selon les termes de la licence Creative Commons “attribution – pas d’utilisation commerciale – partage dans les mêmes conditions” 4.0 internationale.

Régression linéaire	14
Remarque	15
Espace de Hilbert des variables aléatoires de carré intégrable	15
Exercices	17
Un exercice tout bête	17
Mélanges de lois	17
Lois conjuguées	18
Randomisation	18
Etats cachés — indépendance conditionnelle	19
Covariance totale	19
Non-réponse	20
Solutions	21
Un exercice tout bête	21
Mélanges de lois	21
Lois conjuguées	22
Randomisation	25
Etats cachés — indépendance conditionnelle	26
Covariance totale	27
Non-réponse	27
Références	29

On s'est consacré jusqu'à présent à l'étude de (suites de) variables aléatoires indépendantes. En pratique cependant, on rencontre souvent des variables dépendant les unes des autres. Dans le cas de la météo, les variables température, vitesse du vent et pression en fournissent un exemple. Dans les approches bayésiennes, on résume l'information disponible sur l'état du système étudié par la **loi a priori** et on met à jour notre connaissance du système en incorporant de l'information supplémentaire (par exemple des observations). On cherche alors à caractériser la **loi a posteriori** de l'état du système, qui est la loi de l'état sachant l'information supplémentaire. On va ainsi s'attacher dans ce chapitre à décrire les **lois conditionnelles** qui vont permettre de résumer l'information apportée par une variable (ou un vecteur) sur une autre et s'intéresser en particulier à l'**espérance conditionnelle** qui indiquera le comportement moyen d'une variable conditionnellement à une autre. Ce dernier cas pose le cadre probabiliste d'un des problèmes fondamentaux en apprentissage statistique : l'apprentissage supervisé, où on dispose d'un ensemble de réalisations d'une variable dont on cherche à prédire le comportement à partir d'un ensemble de variables dites explicatives (ou prédicteurs).

Lois conditionnelles dans un couple

Soient deux variables aléatoires X et Y définies sur le même espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$. Dans le cas où X et Y sont indépendantes, on a vu que pour tous boréliens B_1 et B_2 de \mathbb{R} , on a

$$\mathbb{P}(X \in B_1, Y \in B_2) = \mathbb{P}(X \in B_1)\mathbb{P}(Y \in B_2) = \mathbb{P}_X(B_1)\mathbb{P}_Y(B_2) = \int_{B_1} \mathbb{P}_Y(B_2)\mathbb{P}_X(dx),$$

où on a utilisé le théorème de Fubini (les mesures de probabilités sont finies, donc σ -finies).

Du fait de l'indépendance, on a aussi $\mathbb{P}_Y(B_2) = \mathbb{P}(Y \in B_2) = \mathbb{P}(Y \in B_2 | X \in B_1) = \mathbb{P}_Y(B_2 | X \in B_1)$ ce qui exprime que pour tout borélien B_1 , la loi conditionnelle de Y sachant $X \in B_1$ est identique à la loi de Y .

Dans le cas général, on va chercher une égalité de la forme

$$\mathbb{P}(X \in B_1, Y \in B_2) = \mathbb{P}_X(B_1)\mathbb{P}_Y(B_2 | X \in B_1) = \int_{B_1} \mathbb{P}_{Y|X=x}(B_2)\mathbb{P}_X(dx)$$

et s'intéresser à caractériser la *loi conditionnelle de Y sachant $X = x$* , que l'on notera donc $\mathbb{P}_{Y|X=x}$.

De même, pour toute application $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ borélienne telle que $g(X, Y)$ admette une espérance (relativement à la loi du couple $\mathbb{P}_{X,Y}$), on voudrait écrire :

$$\mathbb{E}(g(X, Y)) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x, y)\mathbb{P}_{Y|X=x}(dy) \right) \mathbb{P}_X(dx)$$

Pour bien fixer les idées, on va décrire spécifiquement les cas où X est discrète puis où le couple (X, Y) admet une densité avant d'aborder le cas général.

Cas où X est discrète

Dans ce paragraphe, on suppose que la variable aléatoire réelle X est discrète, c'est-à-dire que l'ensemble $X(\Omega) \subset \mathbb{R}$ des valeurs x_k prises par X est au plus dénombrable.

On peut imposer que $\forall x \in X(\Omega)$ on ait $\mathbb{P}(X = x) > 0$, quitte à modifier X sur un ensemble de probabilité nulle. On va ainsi pouvoir utiliser la définition de la probabilité conditionnelle pour des événements de la forme $\{X = x\}$. Ceci

permet d'écrire pour tous boréliens B_1 et B_2 de \mathbb{R} :

$$\begin{aligned}\mathbb{P}(X \in B_1, Y \in B_2) &= \sum_{x \in X(\Omega) \cap B_1} \mathbb{P}(X = x, Y \in B_2) \\ &= \sum_{x \in X(\Omega) \cap B_1} \mathbb{P}(X = x) \mathbb{P}(Y \in B_2 | X = x) \\ &= \int_{B_1} \mathbb{P}(Y \in B_2 | X = x) \mathbb{P}_X(dx)\end{aligned}$$

puisque $\mathbb{P}_X = \sum_{x \in X(\Omega)} \mathbb{P}(X = x) \delta_x$. On obtient ainsi l'écriture souhaitée en posant

$$\mathbb{P}_{Y|X=x}(B_2) = \mathbb{P}(Y \in B_2 | X = x), \quad \forall x \in X(\Omega), \forall B_2 \in \mathcal{B}(\mathbb{R}).$$

Remarque

$\mathbb{P}_{Y|X=x}$ ainsi définie est simplement la probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ image par Y de la probabilité conditionnelle $\mathbb{P}(\cdot | X = x)$ définie sur (Ω, \mathcal{A}) , autrement dit, la **loi de Y relative à $\mathbb{P}(\cdot | X = x)$** et non à \mathbb{P} .

La formule ci-dessus s'écrit $\mathbb{P}_{X,Y}(B_1 \times B_2) = \int_{B_1} \mathbb{P}(Y \in B_2 | X = x) \mathbb{P}_X(dx)$, où $\mathbb{P}_{X,Y}$ est la loi du couple. Elle se généralise à tout borélien B de \mathbb{R}^2 de la manière suivante :

$$\begin{aligned}\mathbb{P}_{X,Y}(B) &= \mathbb{P}((X, Y) \in B) = \sum_{x \in X(\Omega)} \mathbb{P}(X = x, (x, Y) \in B) \\ &= \sum_{x \in X(\Omega)} \mathbb{P}(X = x) \mathbb{P}((x, Y) \in B | X = x) \\ &= \sum_{x \in X(\Omega)} \mathbb{P}(X = x) \mathbb{P}_{Y|X=x}(B_x),\end{aligned}$$

où $B_x = \{y \in \mathbb{R}, (x, y) \in B\}$. Ainsi, pour tout B borélien de \mathbb{R}^2 ,

$$\mathbb{E}(1_B(X, Y)) = \int_{\mathbb{R}^2} 1_B(x, y) \mathbb{P}_{X,Y}(dxdy) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} 1_B(x, y) \mathbb{P}_{Y|X=x}(dy) \right) \mathbb{P}_X(dx)$$

Par linéarité de l'espérance, on peut ainsi exprimer l'espérance d'une fonction étagée. Pour avoir le résultat pour une fonction borélienne positive, on exprime celle-ci comme limite simple d'une suite croissante de fonctions étagées, et on applique le théorème de convergence monotone. Enfin, on applique cette construction à g_+ et g_- pour une fonction g de signe quelconque $\mathbb{P}_{X,Y}$ -intégrable.

En d'autres termes, on reprend le procédé de construction de l'intégrale de Lebesgue. On obtient ainsi la formule souhaitée :

$$\mathbb{E}(g(X, Y)) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x, y) \mathbb{P}_{Y|X=x}(dy) \right) \mathbb{P}_X(dx).$$

Exemple

Soit $X \geq 0$ une variable aléatoire à valeurs dans \mathbb{N} et Y une variable aléatoire réelle positive telle que la loi du couple $\mathbb{P}_{X,Y}$ vérifie pour tout $n \in \mathbb{N}$ et tout borélien B_2 de \mathbb{R} :

$$\mathbb{P}_{X,Y}(\{n\} \times B_2) = (1 - \alpha) \alpha^n \int_{B_2 \cap \mathbb{R}_+^*} e^{-t} \frac{t^n}{n!} dt, \quad 0 < \alpha < 1$$

$\mathbb{P}_{X,Y}$ est bien une probabilité sur \mathbb{R}^2 puisque par convergence monotone :

$$\begin{aligned} \mathbb{P}_{X,Y}(\mathbb{R}^2) &= \mathbb{P}_{X,Y}(\mathbb{N} \times \mathbb{R}) \\ &= \sum_{n \in \mathbb{N}} \mathbb{P}_{X,Y}(\{n\} \times \mathbb{R}) \\ &= \sum_{n \in \mathbb{N}} (1 - \alpha) \alpha^n \int_{\mathbb{R}_+^*} e^{-t} \frac{t^n}{n!} dt \\ &= (1 - \alpha) \int_{\mathbb{R}_+^*} e^{-t} \sum_{n \in \mathbb{N}} \frac{\alpha t^n}{n!} dt \\ &= (1 - \alpha) \int_{\mathbb{R}_+^*} e^{-(1-\alpha)t} dt = 1 \end{aligned}$$

où on aura reconnu la loi exponentielle de paramètre $(1 - \alpha)$. $\forall n \in \mathbb{N}$,

$$\int_{\mathbb{R}_+^*} e^{-t} \frac{t^n}{n!} dt = \int_{\mathbb{R}_+^*} e^{-t} \frac{t^{(n-1)}}{(n-1)!} dt = \dots = \int_{\mathbb{R}_+^*} e^{-t} dt = 1$$

par intégrations par parties itérées. La loi marginale de X s'écrit donc :

$$\forall n \in \mathbb{N}, \mathbb{P}(X = n) = \mathbb{P}_{X,Y}(\{n\} \times \mathbb{R}_+^*) = (1 - \alpha) \alpha^n,$$

loi géométrique de paramètre $(1 - \alpha)$. On en déduit la loi conditionnelle de Y sachant $X = x$:

$$\mathbb{P}_{Y|X=x}(B_2) = \mathbb{P}(Y \in B_2 | X = x) = \frac{\mathbb{P}_{X,Y}(\{x\} \times B_2)}{\mathbb{P}(X = x)} = \int_{B_2 \cap \mathbb{R}_+^*} e^{-t} \frac{t^x}{x!} dt$$

et $\mathbb{P}_{Y|X=x}$ est la donc la loi gamma de paramètre $(x + 1, 1)$.

Densités conditionnelles

On suppose maintenant que le couple (X, Y) admet une densité $f_{X,Y}$ (par rapport à la mesure de Borel-Lebesgue). On note $f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y) dy$ (respectivement $f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y) dx$) la loi marginale de X (resp. de Y). On s'intéresse à caractériser la densité de la variable Y connaissant la valeur prise par la variable X , c'est la *densité conditionnelle* de Y sachant $\{X = x\}$:

Proposition

La formule suivante définit une densité sur \mathbb{R} , pour tout $x \in \mathbb{R}$ tel que $f_X(x) > 0$.

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

Cette fonction s'appelle la *densité conditionnelle de Y sachant $\{X = x\}$* . La probabilité conditionnelle de Y sachant $\{X = x\}$ s'écrit ainsi $\mathbb{P}_{Y|X=x} = f_{Y|X=x} \ell$, où ℓ représente la mesure de Borel-Lebesgue.

Démonstration La preuve est immédiate puisque $f_{Y|X=x}$ est une fonction positive d'intégrale 1. ■

L'interprétation de cette définition est la suivante : la fonction $f_{Y|X=x}$ est la densité de la "loi conditionnelle de Y sachant que $X = x$ ". Bien sûr, nous avons $\mathbb{P}(X = x) = 0$ puisque X admet une densité, donc la phrase ci-dessus n'a pas réellement de sens, mais elle se justifie heuristiquement ainsi : dx et dy étant de "petits" accroissements des variables x et y et lorsque f et f_X sont continues et strictement positives respectivement en (x, y) et x :

$$\begin{aligned} f_X(x)dx &\approx \mathbb{P}(X \in [x, x + dx]) \\ f_{X,Y}(x, y)dx dy &\approx \mathbb{P}(X \in [x, x + dx], Y \in [y, y + dy]) \end{aligned}$$

Par suite

$$\begin{aligned} f_{Y|X=x}(y)dy &\approx \frac{\mathbb{P}(X \in [x, x + dx], Y \in [y, y + dy])}{\mathbb{P}(X \in [x, x + dx])} \\ &\approx \mathbb{P}(Y \in [y, y + dy] | X \in [x, x + dx]) \end{aligned}$$

On a alors le résultat suivant qui résout le problème posé en introduction :

Proposition

Pour toute fonction $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ telle que $g(X, Y)$ admette une espérance, on a :

$$\mathbb{E}(g(X, Y)) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x, y) f_{Y|X=x}(y) dy \right) f_X(x) dx,$$

dont on déduit, en prenant $g = 1_{B_1 \times B_2}$, que :

$$\mathbb{P}(X \in B_1, Y \in B_2) = \int_{B_1} \left(\int_{B_2} f_{Y|X=x}(y) dy \right) f_X(x) dx.$$

Démonstration On a

$$\begin{aligned} \mathbb{E}(g(X, Y)) &= \int_{\mathbb{R}^2} g(x, y) f_{X,Y}(x, y) dy dx \\ &= \int_{\mathbb{R}^2} g(x, y) f_{Y|X=x}(y) f_X(x) dy dx \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x, y) f_{Y|X=x}(y) dy \right) f_X(x) dx, \end{aligned}$$

les calculs étant licites par application du théorème de Fubini et du fait que l'application $x \mapsto \int_{\mathbb{R}} g(x, y) f_{Y|X=x}(y) dy$ est définie pour $f_X(x) > 0$, soit presque partout relativement à la mesure \mathbb{P}_X \blacksquare

Cas général

On peut établir le résultat suivant, qui complète le théorème de Fubini et le résultat d'existence et d'unicité des mesures produits, et que l'on admettra.

Théorème

Soit un couple (X, Y) de variables aléatoires réelles de loi jointe $\mathbb{P}_{X,Y}$, il existe une famille $(\mathbb{P}_{Y|X=x})_{x \in \mathbb{R}}$ de probabilités sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, unique à une égalité \mathbb{P}_X -presque partout¹, qui vérifie pour tous B_1, B_2 boréliens de \mathbb{R} :

$$\mathbb{P}_{X,Y}(B_1 \times B_2) = \int_{B_1} \left(\int_{B_2} \mathbb{P}_{Y|X=x}(dy) \right) \mathbb{P}_X(dx).$$

Ces probabilités sont appelées *lois conditionnelles* de Y sachant $X = x$. On a de plus pour toute application $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ telle que $g(X, Y)$ admette une espérance :

$$\mathbb{E}(g(X, Y)) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x, y) \mathbb{P}_{Y|X=x}(dy) \right) \mathbb{P}_X(dx).$$

1. c'est-à-dire qu'on peut définir ces probabilités de la manière qu'on souhaite pour les boréliens B tels que $\mathbb{P}_X(B) = 0$.

Remarques

- Ce résultat peut être interprété comme un **théorème de Fubini conditionnel**, dans le sens où il permet une intégration séquentielle, mais ici la mesure de probabilité du couple (X, Y) s'exprime comme un produit de mesures dont l'un des termes dépend de la variable d'intégration de l'autre. En particulier, si on change l'ordre d'intégration, on change les mesures qui interviennent.
- Fréquemment, dans les applications, la famille des lois conditionnelles est une donnée du modèle considéré, et leur existence ne pose donc pas de problème !
- On retrouve les cas vus précédemment en notant que pour tout borélien B_1 de \mathbb{R} on a $\mathbb{P}_X(B_1) = \int_{B_1} \mathbb{P}_X(dx) = \sum_{x \in B_1} \mathbb{P}(X = x)$ lorsque X est discrète, et que pour tous boréliens B_1 et B_2 de \mathbb{R} on a $\mathbb{P}_X(B_1) = \int_{B_1} f_X(x)dx$ et $\mathbb{P}_{X,Y}(B_1 \times B_2) = \int_{B_1 \times B_2} f_{X,Y}(x, y)dx dy$.
- Dans tout ce qui précède, les rôles de X et Y peuvent évidemment être inversés.

Conséquences

Le théorème précédent a deux conséquences majeures. Il fournit d'une part un moyen efficace d'identifier la loi marginale de Y connaissant la loi marginale de X et la loi de Y sachant $X = x$. En effet, en notant que pour tout borélien B de \mathbb{R} , $\mathbb{P}_Y(B) = \mathbb{P}_{X,Y}(\mathbb{R} \times B)$ et en appliquant ce théorème, on a la proposition suivante :

Proposition

- La loi marginale \mathbb{P}_Y de Y s'exprime comme la moyenne des lois conditionnelles $\mathbb{P}_{Y|X=x}$ pondérée par la loi de X . Pour tout B borélien de \mathbb{R}

$$\mathbb{P}_Y(B) = \int_{\mathbb{R}} \left(\int_B \mathbb{P}_{Y|X=x}(dy) \right) \mathbb{P}_X(dx) = \int_{\mathbb{R}} \mathbb{P}_{Y|X=x}(B) \mathbb{P}_X(dx)$$

- Dans le cas où X est discrète (à valeurs dans I dénombrable), on retrouve une expression de la formule des probabilités totales et composées :

$$\mathbb{P}_Y(B) = \mathbb{P}(Y \in B) = \sum_{x \in I} \mathbb{P}(Y \in B | X = x) \mathbb{P}(X = x)$$

- Dans le cas où le couple (X, Y) admet une densité, puisqu'on a $f_{X,Y}(x, y) = f_{Y|X=x}(y)f_X(x)$, on obtient l'expression suivante pour la densité marginale :

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y)dx = \int_{\mathbb{R}} f_{Y|X=x}(y)f_X(x)dx.$$

On a en particulier la *formule de Bayes pour les densités* : pour tout x tel que $f_X(x) > 0$ et tout y tel que $f_Y(y) > 0$:

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{Y|X=x}(y)f_X(x)}{f_Y(y)}.$$

Exemple

Poursuivons l'exemple vu plus haut. On rappelle qu'on a déjà identifié la loi marginale de X ainsi que la loi conditionnelle de Y sachant $X = n$ pour $n \in \mathbb{N}$ que l'on rappelle ici :

$$\mathbb{P}(X = n) = (1 - \alpha)\alpha^n, \quad n \in \mathbb{N} \text{ et } \forall B \in \mathcal{B}(\mathbb{R}), \quad \mathbb{P}_{Y|X=n}(B) = \int_{B \cap \mathbb{R}_+^*} e^{-t} \frac{t^n}{n!} dt$$

On peut en déduire la loi marginale de Y en utilisant la proposition précédente et le théorème de convergence monotone :

$$\begin{aligned} \mathbb{P}_Y(B) &= \sum_{n \in \mathbb{N}} (1 - \alpha)\alpha^n \int_{B \cap \mathbb{R}_+^*} e^{-t} \frac{t^n}{n!} dt \\ &= (1 - \alpha) \int_{B \cap \mathbb{R}_+^*} e^{-t} \sum_{n \in \mathbb{N}} \frac{(\alpha t)^n}{n!} dt \\ &= \int_B 1_{\mathbb{R}_+}(t) (1 - \alpha) e^{-(1-\alpha)t} dt, \end{aligned}$$

de sorte que Y suit une loi exponentielle de paramètre $(1 - \alpha)$.

En inversant les rôles, on va pouvoir identifier la loi de X sachant $Y \in B$ en notant que

$$\begin{aligned} \mathbb{P}_{X,Y}(\{n\} \times B) &= \mathbb{P}_X(\{n\}) \mathbb{P}_{Y|X=n}(B) \\ &= \int_B \frac{(\alpha t)^n}{n!} e^{-\alpha t} \mathbb{P}_Y(dt) \\ &= \int_B \mathbb{P}_{X|Y=t}(\{n\}) \mathbb{P}_Y(dt) \end{aligned}$$

où l'on reconnaît que $\mathbb{P}_{X=n|Y=t}(\{n\}) = \frac{(\alpha t)^n}{n!} e^{-\alpha t}$, c'est-à-dire que X sachant $Y = t$ suit une loi de Poisson de paramètre αt pour \mathbb{P}_Y -presque tout t .

En utilisant, le théorème précédent, on obtient également une nouvelle caractérisation de l'indépendance de deux variables aléatoires faisant intervenir les lois conditionnelles.

Proposition (critère d'indépendance)

1. X et Y sont indépendantes si et seulement si, pour \mathbb{P}_X -presque tout x , $\mathbb{P}_{Y|X=x}$ ne dépend pas de x et dans ce cas, on a $\mathbb{P}_{Y|X=x} = \mathbb{P}_Y$, c'est-à-dire que la loi conditionnelle est identique à la loi marginale.
2. Dans le cas où (X, Y) admet une densité, X et Y sont indépendantes si et seulement si la densité conditionnelle de Y sachant $\{X = x\}$ ne dépend pas de x .

Démonstration

1. Si X et Y sont indépendantes, pour tous B_1, B_2 boréliens de \mathbb{R} , $\mathbb{P}_{X,Y}(B_1 \times B_2) = \mathbb{P}_X(B_1)\mathbb{P}_Y(B_2) = \int_{B_1} \mathbb{P}_Y(B_2)\mathbb{P}_X(dx) = \int_{B_2} \mathbb{P}_X(B_1)\mathbb{P}_Y(dy)$. Le résultat d'unicité du théorème de Fubini conditionnel (à une égalité \mathbb{P}_X -presque sûre près), nous indique alors que $\mathbb{P}_{Y|X=x}(B_2) = \mathbb{P}_Y(B_2)$. Inversement, si $\mathbb{P}_{Y|X=x} = \mathbb{P}_Y$, alors $\mathbb{P}_{X,Y}(B_1 \times B_2) = \int_{B_1} \mathbb{P}_{Y|X=x}(B_2)\mathbb{P}_X(dx) = \int_{B_1} \mathbb{P}_Y(B_2)\mathbb{P}_X(dx) = \mathbb{P}_X(B_1)\mathbb{P}_Y(B_2)$.
2. Si X et Y sont indépendantes, $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, d'où $f_{Y|X=x}(y) = f_Y(y)$. Inversement, si $f_{Y|X=x}(y) = f_Y(y)$ alors $f_{X,Y}(x, y) = f_{Y|X=x}(y)f_X(x) = f_Y(y)f_X(x)$ et X et Y sont indépendantes.

■

Espérance conditionnelle

Puisque $\mathbb{P}_{Y|X=x}$ est la loi d'une variable aléatoire, on peut définir l'espérance qui lui est associée et introduire la notion d'espérance conditionnelle dans le cas où Y est intégrable.

Définition

Soit Y une variable aléatoire intégrable.

1. L'espérance conditionnelle de Y sachant $\{X = x\}$ est définie par

$$\mathbb{E}(Y|X = x) = \int_{\mathbb{R}} y \mathbb{P}_{Y|X=x}(dy).$$

2. L'espérance conditionnelle de Y sachant X est la **variable aléatoire** définie par :

$$\mathbb{E}(Y|X) = \psi(X), \text{ avec } \psi(x) = \mathbb{E}(Y|X = x).$$

Remarques

1. $\psi(x)$ n'est définie que pour $x \notin N$, avec $\mathbb{P}(X \in N) = 0$. Par conséquent, la définition définit bien l'espérance conditionnelle $\psi(X) = \mathbb{E}(Y|X)$ \mathbb{P}_X -presque partout, autrement dit avec probabilité 1, ou encore presque sûrement.
2. $\mathbb{E}(\mathbb{E}(|Y||X)) = \mathbb{E}(|Y|)$ comme conséquence directe du théorème de Fubini conditionnel. L'espérance conditionnelle de Y sachant X est bien définie dès que Y est intégrable.
3. Lorsque (X, Y) admet une densité, l'espérance conditionnelle de Y sachant $\{X = x\}$ s'écrit

$$\mathbb{E}(Y|X = x) = \int_{\mathbb{R}} y f_{Y|X=x}(y) dy.$$

On peut étendre cette définition aux variables de la forme $g(X, Y)$.

Définition

Soit Y une variable aléatoire et g une fonction mesurable positive ou $\mathbb{P}_{X,Y}$ -intégrable sur \mathbb{R}^2 .

1. L'espérance conditionnelle de $g(X, Y)$ sachant $\{X = x\}$ est définie par

$$\mathbb{E}(g(X, Y)|X = x) = \int_{\mathbb{R}} g(x, y) \mathbb{P}_{Y|X=x}(dy).$$

2. L'espérance conditionnelle de $g(X, Y)$ sachant X est la **variable aléatoire** définie par :

$$\mathbb{E}(g(X, Y)|X) = \psi(X), \text{ avec } \psi(x) = \mathbb{E}(g(X, Y)|X = x).$$

Théorème

Si Y est intégrable, alors $\psi(X) = \mathbb{E}(Y|X)$ est intégrable, et

$$\mathbb{E}(\psi(X)) = E(Y).$$

Démonstration C'est une conséquence directe du théorème de Fubini conditionnel. ■

Ce résultat permet de calculer $\mathbb{E}(Y)$ en conditionnant par une variable auxiliaire X :

$$\mathbb{E}(Y) = \int_{\mathbb{R}} \mathbb{E}(Y|X = x) \mathbb{P}_X(dx)$$

Il généralise la formule des probabilités totales, qui correspond ici à $Y = 1_A$, et $B_x = \{X = x\}$ où les B_x forment cette fois une partition non dénombrable de \mathbb{R} . On l'écrit souvent sous la forme

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$$

et on l'appelle la *formule de l'espérance totale*.

L'espérance conditionnelle étant définie comme l'espérance de la loi conditionnelle, elle hérite des propriétés usuelles de l'espérance :

1. si Y et Z sont intégrables, $\mathbb{E}(aY + bZ|X) = a\mathbb{E}(Y|X) + b\mathbb{E}(Z|X)$,
2. $\mathbb{E}(Y|X) \geq 0$ si $Y \geq 0$,
3. $\mathbb{E}(1|X) = 1$.

De plus, si g est mesurable positive ou \mathbb{P}_X -intégrable,

$$\mathbb{E}(Yg(X)|X) = g(X)\mathbb{E}(Y|X)$$

est une généralisation de l'égalité 1. ci-dessus, au cas où $a = g(X)$, qui doit être considéré "comme une constante" dans le calcul de l'espérance conditionnelle sachant X (X est fixée comme une donnée connue a priori). En effet, on a alors $\mathbb{E}(g(X)Y|X = x) = g(x)\psi(x)$. Enfin, on déduit directement du théorème de Fubini conditionnel la proposition suivante.

Proposition — transfert conditionnel

Soient un couple (X, Y) de variables aléatoires réelles de loi jointe $\mathbb{P}_{X,Y}$ et g une fonction mesurable positive ou $\mathbb{P}_{X,Y}$ -intégrable sur \mathbb{R}^2 . On a pour \mathbb{P}_X -presque tout x dans \mathbb{R}

$$\mathbb{E}(g(X, Y)|X = x) = \mathbb{E}(g(x, Y)|X = x) = \int_{\mathbb{R}} g(x, y)\mathbb{P}_{Y|X=x}(dy)$$

Si de plus X et Y sont indépendantes, on a :

$$\mathbb{E}(g(X, Y)|X = x) = \mathbb{E}(g(x, Y)|X = x) = \int_{\mathbb{R}} g(x, y)\mathbb{P}_Y(dy).$$

Autrement dit, lorsqu'on conditionne par l'événement $\{X = x\}$, cela revient à fixer la valeur de la variable aléatoire X à la constante x .

Vecteurs Gaussiens à densité

Dans ce qui précède, on a décrit les lois et les espérances conditionnelles dans le cas d'un couple de variables aléatoires à valeurs dans \mathbb{R}^2 . Ces résultats sont aussi valables pour des couples de vecteurs, dont on décrit ici un cas particulier.

Dans le cas des vecteurs gaussiens à densité, c'est-à-dire dont la matrice de covariance est définie positive et donc inversible, le calcul des lois conditionnelles de certaines composantes par rapport aux autres est particulièrement aisé. On va voir en particulier que les lois conditionnelles ont le bon goût d'être elles-mêmes gaussiennes, ce qui explique (en partie) le succès de ces modèles dans les applications.

On considère un vecteur gaussien $X = (X_1, \dots, X_n)$ à valeurs dans \mathbb{R}^n d'espérance m et de matrice de covariance C définie positive. On a vu au chapitre 2 que la densité du vecteur X s'écrit pour $x \in \mathbb{R}^d$:

$$f_X(x) = \frac{1}{(2\pi)^{n/2} \sqrt{\det(C)}} \exp\left(-\frac{1}{2}(x-m)^t C^{-1}(x-m)\right)$$

Soit $1 \leq k < n$ un entier. On souhaite exprimer $f_{Y|Z=z}$, la densité conditionnelle de $Y = (X_1, \dots, X_{k-1})$ sachant $Z = (X_k, \dots, X_n) = (x_k, \dots, x_n) = z$. On a vu que

$$f_X = f_{Y|Z=z} f_Z,$$

où f_Z est la densité marginale de Z . On cherche donc à décomposer f_X de la sorte. On note $m = (m_Y, m_Z)$ et on remarque que C peut se décomposer en blocs :

$$C = \begin{pmatrix} C_Y & C_{Y,Z} \\ C_{Z,Y} & C_Z \end{pmatrix}$$

où $C_Y = \text{Cov}(Y, Y)$, $C_Z = \text{Cov}(Z, Z)$ et $C_{Y,Z} = \text{Cov}(Y, Z)$. Le *complément de Schur*² du bloc C_Y est la matrice

$$CS_Y = C_Y - C_{Y,Z} C_Z^{-1} C_{Z,Y}$$

et permet d'exprimer l'inverse de C comme :

$$C^{-1} = \begin{pmatrix} CS_Y^{-1} & -CS_Y^{-1} C_{Y,Z} C_Z^{-1} \\ -C_Z^{-1} C_{Z,Y} CS_Y^{-1} & C_Z^{-1} + C_Z^{-1} C_{Z,Y} CS_Y^{-1} C_{Y,Z} C_Z^{-1} \end{pmatrix}$$

On peut alors réarranger les termes de la forme quadratique dans f_X et on obtient :

$$\begin{aligned} (x-m)^t C^{-1} (x-m) &= (y - (m_Y + C_{Y,Z} C_Z^{-1} (z - m_Z)))^t CS_Y^{-1} \\ &\quad \cdot (y - (m_Y + C_{Y,Z} C_Z^{-1} (z - m_Z))) \\ &\quad + (z - m_Z)^t C_Z^{-1} (z - m_Z) \end{aligned}$$

Pour la constante, on peut remarquer que :

$$\det(C) = \det(CS_Y) \det(C_Z).$$

2. voir par exemple l'excellent matrix cookbook.

On en déduit ainsi que

$$f_{Y|Z=z}(y) = \frac{1}{(2\pi)^{k/2} \sqrt{\det(CS_Y)}} \exp \left(-\frac{1}{2} (y - \psi(z))^t CS_Y^{-1} (y - \psi(z)) \right)$$

C'est-à-dire que la variable aléatoire $Y|Z = z$ est gaussienne d'espérance $m_{Y|Z=z} = \psi(z) = m_Y + C_{Y,Z} C_Z^{-1} (z - m_Z)$ et de matrice de covariance $CS_Y = C_Y - C_{Y,Z} C_Z^{-1} C_{Z,Y}$. Autrement dit, l'espérance conditionnelle de Y sachant Z est la variable aléatoire $\mathbb{E}(Y|Z) = \psi(Z) = (m_Y + C_{Y,Z} C_Z^{-1} (Z - m_Z))$. On notera que la covariance conditionnelle donnée par CS_Y ne dépend pas de la valeur prise par Z .

Régression et espérance conditionnelle des variables de carré intégrable

La régression est un ensemble de méthodes (d'apprentissage) statistiques très utilisées pour analyser la relation d'une variable par rapport à une ou plusieurs autres. Ces méthodes visent notamment à décrire les liens de dépendance entre variables mais aussi de prédire au mieux la valeur d'une quantité non observée en fonction d'une ou plusieurs autres variables. On va en décrire ici le principe du point de vue probabiliste dans le cas particulier des variables de carré intégrable (ou dans \mathcal{L}^2). On verra dans ce cadre, que l'on rencontre très fréquemment en pratique, une interprétation géométrique très éclairante de l'espérance conditionnelle.

Régression linéaire

On considère deux variables aléatoires réelles, de carré intégrable, définies sur le même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, et dont on suppose connues les variances et la covariance. Nous souhaitons trouver la meilleure approximation de Y par une fonction affine de X de la forme $aX + b$, au sens des moindres carrés, c'est-à-dire qui minimise la quantité $\mathbb{E}((Y - (aX + b))^2)$. Il s'agit de déterminer les constantes a et b telles que $\mathbb{E}((Y - (aX + b))^2)$ soit minimale. Or, par linéarité,

$$\mathbb{E}((Y - (aX + b))^2) = \mathbb{E}(Y^2) - 2a\mathbb{E}(XY) - 2b\mathbb{E}(Y) + a^2\mathbb{E}(X^2) + 2ab\mathbb{E}(X) + b^2.$$

L'annulation de ses dérivées partielles en a et b entraîne que les solutions sont

$$\begin{aligned} a &= \frac{\text{Cov}(X, Y)}{\mathbb{V}(X)} = \rho(X, Y) \frac{\sigma_Y}{\sigma_X} \\ b &= \mathbb{E}(Y) - a\mathbb{E}(X) \end{aligned}$$

On vérifie aisément que ces valeurs donnent bien un minimum pour $\mathbb{E}((Y - (aX + b))^2)$ qui est convexe, et déterminent ainsi la meilleure approximation linéaire de Y basée sur X au sens de l'erreur quadratique moyenne.

Cette approximation linéaire vaut

$$\mathbb{E}(Y) + \rho(X, Y) \frac{\sigma_Y}{\sigma_X} (X - \mathbb{E}(X))$$

et l'erreur quadratique moyenne vaut alors

$$\begin{aligned} \mathbb{E} \left(\left(Y - \mathbb{E}(Y) - \rho(X, Y) \frac{\sigma_Y}{\sigma_X} (X - \mathbb{E}(X)) \right)^2 \right) &= \sigma_Y^2 + \rho^2(X, Y) \sigma_Y^2 - 2\rho^2(X, Y) \sigma_Y^2 \\ &= \sigma_Y^2 (1 - \rho^2(X, Y)). \end{aligned}$$

On voit ainsi que cette erreur est proche de 0 lorsque $|\rho(X, Y)| \approx 1$ tandis qu'elle est proche de $\mathbb{V}(Y) = \sigma_Y^2$ lorsque $\rho(X, Y) \approx 0$. On notera au passage qu'on obtient que la meilleure approximation de Y par une constante est son espérance.

Remarque

L'hypothèse d'une relation linéaire est très forte et pas nécessairement toujours adaptée pour expliquer des relations de dépendances entre variables. Soit en effet une variable aléatoire réelle X de \mathcal{L}^3 symétrique, c'est-à-dire telle que X et $-X$ sont de même loi. On a alors $\mathbb{E}(X) = -\mathbb{E}(X) = 0$. Les variables X et X^2 ne sont clairement pas indépendantes. Pour autant, on a $\text{Cov}(X, X^2) = \mathbb{E}(X^3) = -\mathbb{E}(X^3) = 0$ et le coefficient de régression a ci-dessus est nul.

Espace de Hilbert des variables aléatoires de carré intégrable

Dans le paragraphe précédent, on s'est intéressé à approximer linéairement une variable aléatoire Y de carré intégrable par une autre variable X également de carré intégrable. On va montrer ici que la meilleure approximation, au sens de l'erreur quadratique moyenne, de Y par une fonction de X est précisément donnée par $\psi(X) = \mathbb{E}(Y|X)$. Ce paragraphe fait appel à des notions hors programme et est par conséquent non exigible. Il fournit néanmoins une interprétation géométrique particulièrement frappante de l'espérance conditionnelle.

On a besoin en pratique de travailler sur un espace un peu plus petit que \mathcal{L}^2 tout entier. En effet, les outils que nous allons utiliser ne nous permettent pas de distinguer entre deux variables X et Y égales presque sûrement, c'est-à-dire telles que $\exists N \in \mathcal{A}$, tel que $\mathbb{P}(N) = 0$ et $\forall \omega \in N^c$, $X(\omega) = Y(\omega)$. Cette notion d'égalité presque sûre est une relation d'équivalence. On va ainsi travailler avec l'espace L^2 des classes de variables pour l'égalité presque sûre, c'est-à-dire que

L^2 contiendra un unique représentant de chacune de ces classes. Dans ce cadre, au lieu d'écrire $X = 0$ p.s., on écrit simplement $X = 0$.

On peut d'abord montrer que l'espace vectoriel L^2 des variables aléatoires de carré intégrable forme un espace de Hilbert si on le munit du produit scalaire :

$$\langle X, Y \rangle = \mathbb{E}(XY) \text{ et de la norme associée } \|X\| = \mathbb{E}(X^2)^{1/2}.$$

L'écart-type est ainsi la norme des variables centrées et la covariance le produit scalaire des variables centrées.

Ce produit scalaire est bien défini pour tout couple (X, Y) de variables de L^2 puisque par l'inégalité de Cauchy-Schwartz :

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

et on a bien $\|X\| = 0$ si et seulement si $X = 0$. On peut enfin montrer que L^2 est complet pour la norme définie ci-dessus (voir Jacod and Protter (2003) pour la démonstration).

Soient maintenant X et $Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$. On considère L_X^2 le sous-espace de L^2 constitué des (classes d'équivalence) des variables aléatoires fonctions seulement de X du type $\phi(X)$ (avec ϕ telle que $\phi(X) \in L^2$). On peut montrer que L_X^2 est convexe et fermé.

Alors, l'espérance conditionnelle de Y sachant X , $\mathbb{E}(Y|X)$ s'interprète comme **la projection orthogonale** de Y sur L_X^2 .

Soit en effet l'opérateur qui à $Y \in L^2$ associe $\mathbb{E}(Y|X) \in L_X^2$. On a vu que c'est un opérateur linéaire. Pour montrer qu'il s'agit d'un projecteur orthogonal, on peut vérifier qu'il est idempotent et auto-adjoint :

- on a bien $\mathbb{E}(\mathbb{E}(Y|X)|X) = \mathbb{E}(Y|X)$
- et pour $Z \in L^2$, $\langle Z, \mathbb{E}(Y|X) \rangle = \mathbb{E}(Z\mathbb{E}(Y|X)) = \mathbb{E}(\mathbb{E}(Z|X)\mathbb{E}(Y|X)) = \mathbb{E}(\mathbb{E}(Z|X)\mathbb{E}(Y)) = \langle \mathbb{E}(Z|X), Y \rangle$.

Le théorème de projection sur un convexe fermé dans les espaces de Hilbert³ assure alors que

$$\arg \min_{\phi(X) \in L_X^2} \|Y - \phi(X)\|^2 = \arg \min_{\phi(X) \in L_X^2} \mathbb{E}((Y - \phi(X))^2) = \mathbb{E}(Y|X) = \psi(X)$$

Ainsi, $\mathbb{E}(Y|X)$ est la meilleure approximation (au sens des moindres carrés) de Y par une fonction de X .

Il est alors immédiat que le "résidu" $Y - \mathbb{E}(Y|X)$ est non corrélé avec X du fait de l'orthogonalité. On en déduit la *formule de la variance totale* :

3. voir par exemple les Rappels mathématiques pour la mécanique quantique de Bruno Figliuzzi

$$\begin{aligned}
\mathbb{V}(Y) &= \|Y - \mathbb{E}(Y)\|^2 = \|Y - \mathbb{E}(Y|X) + \mathbb{E}(Y|X) - \mathbb{E}(Y)\|^2 \\
&= \|Y - \mathbb{E}(Y|X)\|^2 + \|\mathbb{E}(Y|X) - \mathbb{E}(Y)\|^2 \\
&= \mathbb{E}((Y - \mathbb{E}(Y|X))^2) + \mathbb{E}((\mathbb{E}(Y|X) - \mathbb{E}(Y))^2) \\
&= \mathbb{E}(\mathbb{E}((Y - \mathbb{E}(Y|X))^2|X)) + \mathbb{V}(\mathbb{E}(Y|X)) \\
&= \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X)).
\end{aligned}$$

où on a utilisé la formule de l'espérance totale et introduit la variable aléatoire variance conditionnelle $\mathbb{V}(Y|X) = \mathbb{E}((Y - \mathbb{E}(Y|X))^2|X)$ comme cas particulier de la définition vue plus haut.

Exercices

Un exercice tout bête

Soient X et Y de densité jointe $f_{X,Y}(x,y) = \frac{1}{x}1_T(x,y)$ où T est le triangle $T = \{0 < y < x < 1\}$.

1. Calculer la densité marginale de X
2. Calculer la densité conditionnelle de Y sachant $X = x$.
3. En déduire l'espérance conditionnelle de Y sachant X .

(?)

Mélanges de lois

Adapté du cours de probabilités de S. Bonnabel et M. Schmidt (MINES Paris-Tech).

Pour modéliser un phénomène multimodal, on utilise souvent des mélanges de gaussiennes. C'est le cas notamment en classification non-supervisée, où on fait l'hypothèse que chacune des classes suit une loi gaussienne. Soient $n \in \mathbb{N}^*$ et K une variable aléatoire prenant les valeurs $1, \dots, n$ avec les probabilités non nulles p_1, \dots, p_n telles que $\sum_{i=1}^n p_i = 1$. Soient X_1, \dots, X_n des variables aléatoires gaussiennes mutuellement indépendantes, d'espérances respectives $m_1, \dots, m_n \in \mathbb{R}$ et de variances respectives $\sigma_1^2, \dots, \sigma_n^2 \in \mathbb{R}_+^*$, toutes indépendantes de K . On appelle mélange de gaussiennes la loi de la variable aléatoire $X = X_K$. Pour tout $i \in \{1, \dots, n\}$, on notera f_i la densité de la variable aléatoire X_i .

Question 1 Soit $i \in \{1, \dots, n\}$. Quelle est la densité $f_{X|K=i}$ de X conditionnellement à l'événement $\{K = i\}$? (?)

Question 2 Calculer la densité de probabilité de la variable X . (?)

Question 3 Calculer $\mathbb{E}(X)$. Montrer que $\mathbb{V}(X) = \sum_{i=1}^n p_i \sigma_i^2 + \bar{\sigma}^2$, où ce dernier terme peut être interprété comme la dispersion des espérances. (?)

Question 4 Comment approximeriez-vous le mélange par une unique gaussienne ? Faire un schéma dans le cas $m = 2$. (?)

Lois conjuguées

Soit un vecteur aléatoire (X, Y) de loi jointe $\mathbb{P}_{X,Y}$. Expliciter la loi conditionnelle de Y sachant $\{X = x\}$ dans les situations suivantes, en prenant soin d'explicitier pour quelles valeurs de x ces dernières ont du sens.

Question 1 Y suit une loi Exponentielle de paramètre $\lambda \in \mathbb{R}_+^*$ et pour tout $y \in \mathbb{R}_+^*$, la variable aléatoire X sachant $\{Y = y\}$ suit une loi Exponentielle de paramètre y . (?)

Question 2 Y suit une loi Gamma de paramètres $\alpha, \theta \in \mathbb{R}_+^*$ et pour tout $y \in \mathbb{R}_+^*$, la variable aléatoire X sachant $\{Y = y\}$ suit une loi de Poisson de paramètre y . (?)

Randomisation

Extrait du cours de probabilités de S. Bonnabel et M. Schmidt (MINES Paris-Tech).

Des clients arrivent à la boutique SNCF du boulevard Saint-Michel à des instants aléatoires. On note T_0 l'heure d'ouverture puis T_1, T_2, \dots les temps successifs d'arrivée des clients jusqu'à l'heure de fermeture. Les études statistiques montrent qu'on peut, dans une tranche horaire donnée, supposer que les temps d'attente $X_1 = T_1 - T_0, X_2 = T_2 - T_1, \dots$ peuvent être modélisés par des variables aléatoires indépendantes et de même loi qu'une variable aléatoire positive X . Par ailleurs, une loterie interne décide que chaque jour dans la tranche horaire considérée, le $N^{\text{ème}}$ client sera l'heureux gagnant d'un trajet gratuit Paris-La Ciotat, où N est une variable aléatoire bornée dont la loi dépend du processus de loterie (e.g. tous les clients entre le premier et le 30^{ème} ont une chance 1/30 d'être tirés au sort, en supposant qu'on est sûr d'avoir au moins 30 clients dans la tranche horaire).

On se demande alors : quel est le temps d'attente moyen avant d'obtenir un gagnant ? (?)

Etats cachés — indépendance conditionnelle

Soucieux de l'évolution du potager de l'école, des élèves à la main verte s'intéressent à l'évolution de la température dans le jardin côté Luxembourg. Ils récupèrent pour cela un thermomètre dans un laboratoire, l'installent près du potager, et en relèvent les mesures à intervalles de temps réguliers. Les résultats les surprennent rapidement : les températures affichées ne correspondent pas à celles prévues par météo-France. Leur thermomètre est sans doute déréglé.

On se propose de les aider à comprendre le phénomène dont ils sont témoins à l'aide d'un modèle probabiliste particulier, nommé *modèle de Markov caché*. Précisément, on considère la suite des vraies températures que l'on aurait souhaité relever comme une suite de v.a.r. non indépendantes $(X_n)_{n \in \mathbb{N}^*}$, dite *d'états cachés* (on ne les observe pas directement). Les erreurs commises par le thermomètre sont quant à elles modélisées par une suite de v.a.r. $(\epsilon_n)_{n \in \mathbb{N}^*}$, toutes indépendantes et de même loi admettant une densité f_ϵ . Elles sont supposées indépendantes de la suite $(X_n)_{n \in \mathbb{N}^*}$ (l'erreur du thermomètre lui est propre et ne dépend pas de la température réelle). A chaque instant $n \in \mathbb{N}^*$, on suppose que la mesure du thermomètre est la variable aléatoire

$$Y_n = X_n + \epsilon_n,$$

et que le vecteur aléatoire (X_1, \dots, X_n) possède une densité jointe notée $f_{1:n}$.

Question 1 Montrer que pour tout $n \in \mathbb{N}^*$ et tout $x \in \mathbb{R}$, la loi de Y_n sachant $\{X_n = x\}$ admet une densité, que l'on explicitera. (?)

Question 2 Montrer que les $n \in \mathbb{N}^*$ relevés de température Y_1, \dots, Y_n sont indépendants **conditionnellement** aux états cachés X_1, \dots, X_n . (?)

Covariance totale

Soient X , Y et Z trois variables aléatoires réelles de carré intégrable. La covariance conditionnelle de X et Y sachant Z est définie comme la variable aléatoire

$$\text{Cov}(X, Y \mid Z) = \mathbb{E}\left((X - \mathbb{E}(X \mid Z))(Y - \mathbb{E}(Y \mid Z)) \mid Z\right).$$

Etablir la formule de la covariance totale :

$$\text{Cov}(X, Y) = \mathbb{E}(\text{Cov}(X, Y \mid Z)) + \text{Cov}(\mathbb{E}(X \mid Z), \mathbb{E}(Y \mid Z)).$$

(?)

Non-réponse

Inspiré du cours de probabilité de M. Christine (ENSAE ParisTech).

Un questionnaire est diffusé aux $n \in \mathbb{N}^*$ étudiants de l'école pour savoir combien de temps ils ont consacré à l'étude des probabilités ce semestre. On note Y_i le temps de travail de l'étudiant $i \in \{1, \dots, n\}$ et X_i la variable valant 1 s'il a répondu au questionnaire et 0 sinon. On suppose que les $(X_1, Y_1), \dots, (X_n, Y_n)$ sont des vecteurs aléatoires indépendants de même distribution qu'un vecteur générique (X, Y) tel que

- X est une variable de Bernoulli de paramètre $p \in]0, 1[$ indiquant la probabilité de réponse,
- Y est positive, de carré intégrable, d'espérance $m \in \mathbb{R}_+$ et de variance $\sigma^2 \in \mathbb{R}_+^*$. Le coefficient de corrélation entre X et Y est enfin noté $\rho \in [-1, 1]$.

Question 1 En reprenant la définition de l'espérance conditionnelle $\mathbb{E}(Y \mid X)$ comme meilleure approximation au sens des moindres carrés de Y par une fonction de X , montrer qu'elle coïncide ici avec l'approximation affine de Y par X puis l'écrire en fonction de m, ρ, σ et p . (?)

Question 2 On pose $m_0 := \mathbb{E}(Y \mid X = 0)$ et $m_1 = \mathbb{E}(Y \mid X = 1)$. Calculer m_0 et m_1 en fonction de m, ρ, σ et p . (?)

Question 3 On pose $\sigma_0^2 := \mathbb{V}(Y \mid X = 0)$ et $\sigma_1^2 := \mathbb{V}(Y \mid X = 1)$. Vérifier l'égalité

$$\sigma^2 = \frac{(1-p)\sigma_0^2 + p\sigma_1^2}{1-\rho^2}.$$

(?)

Question 4 Que dire des résultats obtenus aux questions 2 et 3 lorsque :

- X et Y sont non corrélées,
- X et Y sont indépendantes ?

(?)

Solutions

Un exercice tout bête

La densité marginale de X est donnée par $f_X(x) = \int f_{X,Y}(x,y)dy = 1_{]0,1[}(x)$ et pour $x \in]0,1[$,

$$f_{Y|X=x}(y) = \frac{1}{x} 1_{]0,x[}(y)$$

Ainsi X est uniformément distribué sur $]0,1[$, et la loi de Y sachant $X = x$ est uniforme sur $]0,x[$ pour $(0 < x < 1)$. Pour un tel x , l'espérance conditionnelle $\mathbb{E}(Y|X = x)$ vaut ainsi $x/2$ et nous obtenons $\mathbb{E}(Y|X) = \frac{X}{2}$.

Mélanges de lois

Question 1 Soit B un borélien. Par indépendance de K avec X_i , on a

$$\mathbb{P}(X \in B \mid K = i) = \mathbb{P}(X_i \in B \mid K = i) = \mathbb{P}(X_i \in B).$$

La loi de X sachant $\{K = i\}$ est donc la même que celle de X_i , d'où

$$f_{X|K=i} : x \in \mathbb{R} \mapsto f_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left\{ -\frac{(x - m_i)^2}{2\sigma_i^2} \right\}.$$

Question 2 Soit B un borélien. D'après la formule des probabilités totales et la question précédente, on a

$$\mathbb{P}(X \in B) = \sum_{i=1}^n p_i \mathbb{P}(X \in B \mid K = i) = \sum_{i=1}^n p_i \mathbb{P}(X_i \in B).$$

La variable aléatoire X admet donc une densité, qui vaut

$$f_X : x \in \mathbb{R} \mapsto \sum_{i=1}^n p_i f_i(x).$$

Question 3 D'après la question précédente, X a pour espérance

$$\begin{aligned} \mathbb{E}(X) &= \int_{\mathbb{R}} x f_X(x) dx = \int_{\mathbb{R}} x \sum_{i=1}^n p_i f_i(x) dx = \sum_{i=1}^n p_i \int_{\mathbb{R}} x f_i(x) dx \\ &= \sum_{i=1}^n p_i m_i. \end{aligned}$$

Quant à la variance de X , en utilisant l'égalité $\sum_{i=1}^n p_i = 1$, elle vaut

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \int_{\mathbb{R}} x^2 f_X(x) dx - \left(\sum_{i=1}^n p_i m_i \right)^2 \\ &= \sum_{i=1}^n p_i (\sigma_i^2 + m_i^2) - \sum_{i=1}^n p_i \left(\sum_{j=1}^n p_j m_j \right)^2 \\ &= \sum_{i=1}^n p_i \sigma_i^2 + \sum_{i=1}^n p_i \left(m_i - \sum_{j=1}^n p_j m_j \right)^2.\end{aligned}$$

On retrouve bien la forme désirée, avec la dispersion des espérances

$$\bar{\sigma}^2 := \sum_{i=1}^n p_i \left(m_i - \sum_{j=1}^n p_j m_j \right)^2.$$

Question 4 Si l'on souhaite approcher la loi de X avec une unique Gaussienne, et non un mélange, les questions précédentes suggèrent de prendre celle d'espérance $\sum_{i=1}^n p_i m_i$ et de variance $\sum_{i=1}^n p_i \sigma_i^2 + \bar{\sigma}^2$. Voir figure ci-dessous.

Lois conjuguées

On considère dans tout cet exercice B_1 et B_2 des Boréliens.

Question 1 D'après les hypothèses on a

$$\begin{aligned}\mathbb{P}_{X,Y}(B_1 \times B_2) &= \int_{B_2} \left(\int_{B_1} \mathbb{P}_{X|Y=y}(dx) \right) \mathbb{P}_Y(dy) \quad \text{par théorème,} \\ &= \int_{B_2} \left(\int_{B_1} y e^{-yx} 1_{\mathbb{R}_+^*}(x) dx \right) \lambda e^{-\lambda y} 1_{\mathbb{R}_+^*}(y) dy \\ &= \int_{B_1} \int_{B_2} \lambda y e^{-(x+\lambda)y} 1_{\mathbb{R}_+^*}(x) 1_{\mathbb{R}_+^*}(y) dy dx \quad \text{par Fubini.}\end{aligned}$$

Le vecteur aléatoire (X, Y) possède donc une densité jointe

$$f_{X,Y} : (x, y) \in \mathbb{R}^2 \mapsto \lambda y e^{-(x+\lambda)y} 1_{\mathbb{R}_+^*}(x) 1_{\mathbb{R}_+^*}(y).$$

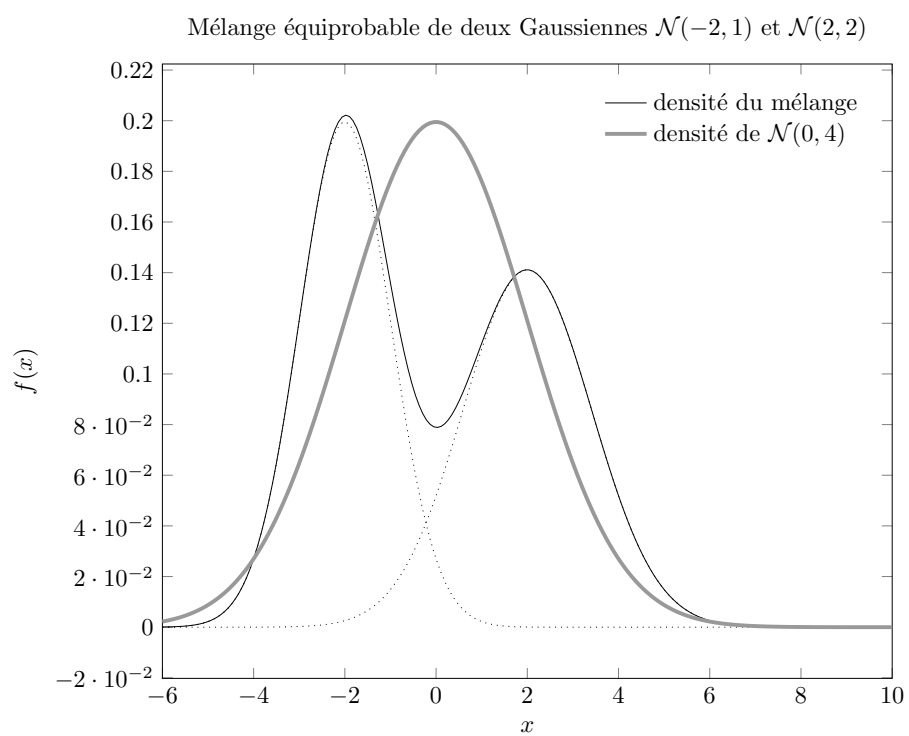


FIGURE 1 – Illustration

La variable aléatoire X a donc aussi une densité : pour tout $x \in \mathbb{R}$

$$\begin{aligned} f_X(x) &= \int_{\mathbb{R}} f_{X,Y}(x,y) dy = \int_{\mathbb{R}} \lambda y e^{-(x+\lambda)y} 1_{\mathbb{R}_+^*}(x) 1_{\mathbb{R}_+^*}(y) dy \\ &= \begin{cases} \frac{\lambda}{x+\lambda} \int_0^{+\infty} y(x+\lambda) e^{-(x+\lambda)y} dy & \text{si } x > 0, \\ 0 & \text{sinon.} \end{cases} \end{aligned}$$

On reconnaît dans cette dernière intégrale la formule de l'espérance d'une loi Exponentielle de paramètre $x+\lambda$, et on en déduit que pour tout $x \in \mathbb{R}$

$$f_X(x) = \frac{\lambda}{(x+\lambda)^2} 1_{\mathbb{R}_+^*}(x).$$

Pour tout $x \in \mathbb{R}_+^*$ la variable Y sachant $\{X=x\}$ admet donc aussi une densité, que l'on explicite avec la formule de Bayes : pour tout $y \in \mathbb{R}$

$$\begin{aligned} f_{Y|X=x}(y) &= \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{\lambda y e^{-(x+\lambda)y} 1_{\mathbb{R}_+^*}(y)}{\frac{\lambda}{(x+\lambda)^2}} \\ &= (x+\lambda)^2 y e^{-(x+\lambda)y} 1_{\mathbb{R}_+^*}(y). \end{aligned}$$

Comme $\Gamma(2) = 1$, on reconnaît ici la densité d'une loi Gamma d'indice 2 et de paramètre d'échelle $x+\lambda$.

Question 2 D'après les hypothèses, en procédant comme précédemment, on a

$$\begin{aligned} \mathbb{P}_{X,Y}(B_1 \times B_2) &= \int_{B_2} \left(\int_{B_1} \mathbb{P}_{X|Y=y}(dx) \right) \mathbb{P}_Y(dy) \\ &= \int_{B_2} \left(\sum_{x \in B_1} \frac{y^x}{x!} e^{-y} 1_{\mathbb{N}}(x) \right) \frac{\theta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\theta y} 1_{\mathbb{R}_+}(y) dy \\ &= \sum_{x \in B_1 \cap \mathbb{N}} \left(\frac{1}{x!} \int_{B_2 \cap \mathbb{R}_+} \frac{\theta^\alpha}{\Gamma(\alpha)} y^{x+\alpha-1} e^{-(\theta+1)y} dy \right) \\ &= \sum_{x \in B_1 \cap \mathbb{N}} \left(\frac{\Gamma(x+\alpha) \theta^\alpha}{x! \Gamma(\alpha) (\theta+1)^{x+\alpha}} \right. \\ &\quad \left. \times \int_{B_2 \cap \mathbb{R}_+} \frac{(\theta+1)^{x+\alpha}}{\Gamma(x+\alpha)} y^{x+\alpha-1} e^{-(\theta+1)y} dy \right). \end{aligned}$$

On reconnaît dans cette dernière intégrale la densité d'une loi Gamma d'indice $x+\alpha$ et de paramètre d'échelle $\theta+1$, qui correspond exactement à la loi conditionnelle de Y sachant $\{X=x\}$ pour $x \in \mathbb{N}$. En effet, on a d'une part

$$\mathbb{P}_X(B_1) = \mathbb{P}_{X,Y}(B_1 \times \mathbb{R}) = \sum_{x \in B_1 \cap \mathbb{N}} \frac{\theta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(x+\alpha)}{x! (\theta+1)^{x+\alpha}},$$

ce qui donne bien pour tout $x \in \mathbb{N}$:

$$\begin{aligned}\mathbb{P}_{Y|X=x}(B_2) &= \mathbb{P}(Y \in B_2 \mid X = x) = \frac{\mathbb{P}_{X,Y}(\{x\} \times B_2)}{\mathbb{P}_X(\{x\})} \\ &= \int_{B_2 \cap \mathbb{R}_+} \frac{(\theta + 1)^{x+\alpha}}{\Gamma(x + \alpha)} y^{x+\alpha-1} e^{-(\theta+1)y} dy.\end{aligned}$$

Randomisation

En termes probabilistes et selon les notations de l'exercice, il s'agit de calculer $\mathbb{E}(T_N - T_0)$, où la variable aléatoire T_N peut s'écrire en fonction d'une somme aléatoire de variables aléatoires indépendantes :

$$T_N = \sum_{i=1}^N X_i + T_0.$$

Comme la boutique ferme au bout d'un certain temps, toutes les variables aléatoires figurant dans l'équation précédente sont bornées, donc intégrables. On peut ainsi calculer $\mathbb{E}(T_N - T_0)$ à l'aide de la formule de l'espérance totale :

$$\mathbb{E}(T_N - T_0) = \mathbb{E}(\mathbb{E}(T_N \mid N)) - T_0.$$

Pour tout $n \in \mathbb{N}^*$ l'énoncé suggère que N est indépendante de X_1, \dots, X_n , elles-mêmes indépendantes et de même loi que X , d'où :

$$\mathbb{E}(T_n \mid N = n) = \sum_{i=1}^n \mathbb{E}(X_i \mid N = n) = \sum_{i=1}^n \mathbb{E}(X_i) = n\mathbb{E}(X).$$

Ainsi, en posant $\psi : n \in \mathbb{N}^* \mapsto n\mathbb{E}(X)$, on obtient

$$\mathbb{E}(T_N - T_0) = \mathbb{E}(\psi(N)) - T_0 = \mathbb{E}(N)\mathbb{E}(X) - T_0.$$

C'était prévisible : en posant arbitrairement $T_0 = 0$, le temps d'attente moyen est le temps d'attente moyen entre deux arrivées, multiplié par le rang moyen du gagnant. Si la loterie dépendait des temps d'arrivées, par exemple en faisant gagner le premier client qui arrive au moins 10 minutes après le client précédent, ψ , et donc le résultat, seraient différents.

Etats cachés — indépendance conditionnelle

Question 1 Soit $n \in \mathbb{N}^*$. Quels que soient $x \in \mathbb{R}$ et B borélien on a

$$\begin{aligned}\mathbb{P}_{Y_n|X_n=x}(B) &= \mathbb{E}(1_B(X_n + \epsilon_n) \mid X_n = x) \\ &= \int_{\mathbb{R}} 1_B(x + y) \mathbb{P}_{\epsilon_n|X_n=x}(dy) \\ &= \int_{\mathbb{R}} 1_B(x + y) f_{\epsilon}(y) dy \quad \text{par indépendance de } X_n \text{ et } \epsilon_n \\ &= \int_B f_{\epsilon}(y - x) dy.\end{aligned}$$

Ainsi, $\mathbb{P}_{Y_n|X_n=x}$ admet bien une densité :

$$f_{Y_n|X_n=x} : y \in \mathbb{R} \mapsto f_{\epsilon}(y - x).$$

Question 2 Soient $n \in \mathbb{N}^*$, $(x_1, \dots, x_n) \in \mathbb{R}^n$ et B_1, \dots, B_n des boréliens. Pour simplifier les écritures, on note $x_{1:n}$ tout vecteur (x_1, \dots, x_n) de \mathbb{R}^n . Alors

$$\begin{aligned}\mathbb{P}_{Y_{1:n}|X_{1:n}=x_{1:n}}(B_1 \times \dots \times B_n) &= \mathbb{E}\left(\prod_{i=1}^n 1_{B_i}(X_i + \epsilon_i) \mid X_{1:n} = x_{1:n}\right) \\ &= \int_{\mathbb{R}^n} \prod_{i=1}^n 1_{B_i}(x_i + y_i) \mathbb{P}_{\epsilon_{1:n}|X_{1:n}=x_{1:n}}(dy_{1:n}) \\ &= \int_{\mathbb{R}^n} \prod_{i=1}^n 1_{B_i}(x_i + y_i) \mathbb{P}_{\epsilon_{1:n}}(dy_{1:n}) \quad \text{par indépendance des } \epsilon_i \text{ et } X_j, \\ &= \prod_{i=1}^n \int_{\mathbb{R}} 1_{B_i}(x_i + y_i) f_{\epsilon}(x_i) dy_i \quad \text{par Fubini et indépendance et même loi des } \epsilon_i, \\ &= \prod_{i=1}^n \int_{\mathbb{R}} 1_{B_i}(y_i) f_{\epsilon}(y_i - x_i) dy_i \\ &= \prod_{i=1}^n \int_{\mathbb{R}} 1_{B_i}(y_i) f_{Y_i|X_i=x_i}(y_i) dy_i \quad \text{par la question 1,} \\ &= \prod_{i=1}^n \mathbb{P}_{Y_i|X_i=x_i}(B_i).\end{aligned}$$

Les n relevés de température sont donc bien indépendants conditionnellement aux états cachés.

Covariance totale

Tout d'abord, par linéarité de l'espérance conditionnelle on a :

$$\begin{aligned}\text{Cov}(X, Y \mid Z) &= \mathbb{E}\left((X - \mathbb{E}(X \mid Z))(Y - \mathbb{E}(Y \mid Z)) \mid Z\right) \\ &= \mathbb{E}\left(XY - X\mathbb{E}(Y \mid Z) - Y\mathbb{E}(X \mid Z) + \mathbb{E}(X \mid Z)\mathbb{E}(Y \mid Z) \mid Z\right) \\ &= \mathbb{E}(XY \mid Z) - \mathbb{E}(X \mid Z)\mathbb{E}(Y \mid Z).\end{aligned}$$

En utilisant la formule de l'espérance totale et la linéarité de l'espérance, on obtient alors

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(\mathbb{E}(XY \mid Z)) - \mathbb{E}(\mathbb{E}(X \mid Z))\mathbb{E}(\mathbb{E}(Y \mid Z)) \\ &= \mathbb{E}(\mathbb{E}(XY \mid Z) - \mathbb{E}(X \mid Z)\mathbb{E}(Y \mid Z)) \\ &\quad + \mathbb{E}(\mathbb{E}(X \mid Z)\mathbb{E}(Y \mid Z)) - \mathbb{E}(\mathbb{E}(X \mid Z))\mathbb{E}(\mathbb{E}(Y \mid Z)) \\ &= \mathbb{E}(\text{Cov}(X, Y \mid Z)) + \text{Cov}(\mathbb{E}(X \mid Z), \mathbb{E}(Y \mid Z)).\end{aligned}$$

Non-réponse

Question 1 L'espérance conditionnelle de Y sachant X peut s'écrire comme la solution au problème de minimisation

$$\min_{\phi(X) \in L_X^2} \mathbb{E}\left((Y - \phi(X))^2\right).$$

Or pour $\phi(X) \in L_X^2$ on a ici

$$\mathbb{E}\left((Y - \phi(X))^2\right) = \mathbb{E}\left((Y - \phi(1))^2 1_{\{1\}}(X)\right) + \mathbb{E}\left((Y - \phi(0))^2 1_{\{0\}}(X)\right),$$

il suffit donc de résoudre pour tout $x \in \{0, 1\}$

$$\min_{\lambda \in \mathbb{R}} \mathbb{E}\left((Y - \lambda)^2 1_{\{x\}}(X)\right).$$

Soit $x \in \{0, 1\}$ et posons $J_x : \lambda \in \mathbb{R} \mapsto \mathbb{E}\left((Y - \lambda)^2 1_{\{x\}}(X)\right)$. Alors pour tout $\lambda \in \mathbb{R}$

$$J_x(\lambda) = \mathbb{E}\left(Y^2 1_{\{x\}}(X)\right) + \lambda^2 \mathbb{P}(X = x) - 2\lambda \mathbb{E}\left(Y 1_{\{x\}}(X)\right)$$

et sa dérivée

$$J'_x(\lambda) = 2\lambda \mathbb{P}(X = x) - 2\mathbb{E}\left(Y 1_{\{x\}}(X)\right)$$

s'annule en

$$\lambda_x := \frac{\mathbb{E}\left(Y 1_{\{x\}}(X)\right)}{\mathbb{P}(X = x)} = \mathbb{E}(Y \mid X = x).$$

On en conclut que

$$\mathbb{E}(Y \mid X) = \mathbb{E}(Y \mid X = 1)1_{\{1\}}(X) + \mathbb{E}(Y \mid X = 0)1_{\{0\}}(X).$$

Or on remarque que $1_{\{1\}}(X) = X$ et $1_{\{0\}}(X) = 1 - X$, ce qui fait de $\mathbb{E}(Y \mid X)$ une fonction affine de X . Elle est par définition la meilleure approximation de Y par une fonction de X , elle coïncide donc avec l'approximation affine de Y par X :

$$\mathbb{E}(Y \mid X) = m + \frac{\rho \sigma}{\sqrt{p(1-p)}} (X - p).$$

Question 2 D'après la question précédente, on a $\mathbb{E}(Y \mid X) = m_0 + (m_1 - m_0)X$, la meilleure approximation affine de Y par X . Ainsi, m_0 et m_1 satisfont

$$\begin{cases} m_1 - m_0 = \frac{\rho \sigma}{\sqrt{p(1-p)}}, \\ m_0 = m - (m_1 - m_0)p, \end{cases} \Leftrightarrow \begin{cases} m_1 = m + \rho \sigma \sqrt{\frac{1-p}{p}}, \\ m_0 = m - \rho \sigma \sqrt{\frac{p}{1-p}}. \end{cases}$$

Question 3 Par la formule de la variance totale et d'après la question 1, on a

$$\begin{aligned} \sigma^2 &= \mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y \mid X)) + \mathbb{V}(\mathbb{E}(Y \mid X)) \\ &= p \sigma_1^2 + (1-p) \sigma_0^2 + \frac{\rho^2 \sigma^2}{p(1-p)} \mathbb{V}(X) \\ &= p \sigma_1^2 + (1-p) \sigma_0^2 + \rho^2 \sigma^2. \end{aligned}$$

Cette égalité se simplifie et donne bien

$$\sigma^2 = \frac{(1-p) \sigma_0^2 + p \sigma_1^2}{1 - \rho^2}.$$

Question 4 Lorsque X et Y sont non corrélées, i.e. $\rho = 0$, on obtient $m_0 = m_1 = m$ puis $\sigma^2 = (1-p) \sigma_0^2 + p \sigma_1^2$. En d'autres termes, $\mathbb{E}(Y \mid X) = m$ est une variable aléatoire constante, et $\mathbb{E}(\mathbb{V}(Y \mid X)) = \sigma^2$. Dans ce cas, la non-réponse n'affecte pas l'espérance, mais potentiellement la variance (la dispersion du temps de travail peut être différente chez les répondants et les non-répondants). Ces deux propriétés sont encore vraies en cas d'indépendance entre X et Y , puisque l'indépendance implique la non corrélation, mais nous avons de plus $\mathbb{V}(Y \mid X) = \sigma^2 = \sigma_1^2 = \sigma_0^2$; la variable aléatoire $\mathbb{V}(Y \mid X)$ est elle aussi constante. Cette fois-ci, la dispersion est la même chez les répondants et les non-répondants : la non-réponse n'affecte pas la variance.

Références

Jacod, J., and P. Protter. 2003. *L'essentiel En Théorie Des Probabilités*. Cassini.
<https://hal.archives-ouvertes.fr/hal-00104956>.