CONTAMINATION DES LOIS (C8)
(09 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'expression **contamination des lois** peut s'interpréter selon deux approches : soit en termes de **lois**, soit en termes de **variables**.

(i) **Contamination d'une loi**. Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé** (cf **espace fondamental**), $(\mathcal{X}, \mathcal{B})$ un **espace mesurable** (eg un **espace d'observation**), $\xi : \Omega \mapsto \mathcal{X}$ et $\varepsilon : \Omega \mapsto \mathcal{X}$ deux **variables aléatoires** à valeurs dans le même **ensemble** $\mathcal{X}$.

(a) on suppose que $\xi \sim P^{\xi}$ et $\varepsilon \sim P^{\varepsilon}$ sont deux **lois de probabilité** définies sur $\mathcal{B}$ et $\lambda \in [0, 1]$ un nombre réel définissant le mélange simple P^{ζ} suivant (cf **mélange de lois**) :

(1) $P^{\zeta} = (1 - \lambda) . P^{\xi} + \lambda . P^{\varepsilon}$.

On dit que P^{ε} **contamine** P^{ξ} , ou qu'il y a **contamination** de P^{ξ} par P^{ε} . On dit que λ est le **taux de contamination** : souvent, $\lambda << 1$;

Un **modèle statistique** dont les lois possibles sont de la forme (1) est parfois appelé **modèle avec erreur grossière**, ou **modèle à erreur grossière**.

(b) dans le même cadre que le précédent, soit $X : \Omega \mapsto \mathcal{X}^N$ un N-**échantillon** de **loi** P^X, image de P par X.

On dit que X est un **échantillon contaminé** ssi il existe une **va** N-dimensionnelle $E : \Omega \mapsto \mathcal{X}^N$, dont la loi est P^E, définissant le mélange simple (ie à deux composantes) P^Z suivant :

(2) $P^Z = (1 - \lambda) . P^X + \lambda . P^E$.

Autrement dit :

(3) $P^Z (B) = P (Z^{-1} (B)) = (1 - \lambda) . P (X^{-1} (B)) + \lambda . P (E^{-1} (B)), \qquad \forall B \in \mathcal{B}$.

On dit que P^E **contamine** P^X ;

(c) plus généralement, la **contamination** d'une loi donnée P^X peut résulter de plusieurs lois (cf **mélange de lois**), eg :

(4) $P^Z = (1 - \lambda) . P^X + \Sigma_{j=2}^{p} \lambda_j . P^{E(j)}$,

avec $(1 - \lambda, \lambda_2 ,..., \lambda_p) \in S_p$ (**simplexe** de $\mathbf{R}^p$), ie $\lambda \geq 0$, $\lambda_j \geq 0$ ($\forall$ j = 2 ,..., p) et $\lambda + \Sigma_{j=2}^{p} \lambda_j = 1$, en notant E(j) pour désigner la variable contaminante E_j (j = 2 ,..., p).

(ii) **Contamination d'une variable**. Les définitions précédentes étaient basées sur des **lois**. On parle aussi de **contamination** lorsqu'une **va** ξ est « altérée » par une autre va ε, eg lorsqu'on observe (forme additive) $\zeta = \xi + \varepsilon$ au lieu d'observer ξ (cf **modèle à erreurs sur les variables**, **modèle de régression**).

Plus généralement, il y a contamination lorsque des va $(\varepsilon_l)_{l=2,...,L}$ altèrent ξ selon (forme additive) :

(4) $\zeta = \xi + \Sigma_{l=2}^{L} \varepsilon_l$.

Dans le cas général, au lieu d'observer ξ, on observe $\zeta = \phi(\xi, \varepsilon_2, ..., \varepsilon_L)$, où ϕ est une fonction « régulière » tq eg :

(5) $\phi(\xi, 0, ..., 0) = \xi$ (ie $\zeta = \xi$ en l'absence d'altération).

La **loi** P^{ζ} de ζ est donc une loi mélangée d'un type particulier, la composante d'intérêt étant P^{ξ} (loi de ξ).

(iii) Comme avec un mélange de lois, une contamination pose divers problèmes :

(a) un **problème d'estimation**. Outre les (éventuels) **paramètres** propres aux lois mises en jeu, on doit aussi estimer les **coefficients (ou poids) du mélange** (eg le coefficient λ dans (1)) ;

(b) un **problème de test**. On peut vouloir tester l'**hypothèse de non contamination** (« **pureté légale** ») $H_0 : \lambda = 0$;

(c) un problème de **robustesse** (estimation ou tests précédents).

Enfin, la contamination constitue un cadre d'analyse important pour l'étude des problèmes d'**aberrations**.