ABERRATION (G9, H4, I5, J5, K13, L6)
(31 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

On est en présence d'**aberration** lorsque, dans une **situation statistique** donnée, les **données** observées semblent engendrées par un autre **schéma probabiliste** que celui « attendu » (« effet de surprise »). Cette notion est cependant une notion relative, notamment lorsqu'un mélange de lois est justifié (par nature, par construction, etc).

(i) L' « anomalie » observée peut prendre des formes variées, traduisant eg une **hétérogénéité** interne à la population concernée, eg notamment :

a) en termes d'**écart** ou de **déviation** pr à une **valeur centrale** ;

(b) en termes de **dispersion** : rupture de **variabilité**, ou variabilité hétérogène.

La terminologie peut varier : une aberration est aussi appelée **observation aberrante**, **valeur aberrante**, **valeur atypique** ou encore **anomalie** (aspects statistiques), ou encore **point aberrant** ou **point atypique** (aspects géométriques ou graphiques).

On dit aussi qu'elle « contamine » l'échantillon observé (cf **contamination des lois**).

Trois causes principales peuvent être à l'origine d'aberrations :

(a) **erreur** d'observation portant **sur les variables** (cf **modèle à erreurs sur les variables**) ;

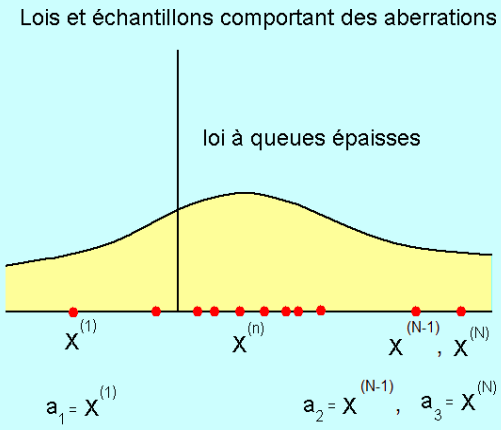
(b) **erreur de spécification** portant sur le modèle (eg famille des lois erronée, liste de variables erronée). Ainsi, certains « **résidus** » d'un **modèle de régression** peuvent paraître aberrants si une variable exogène (pertinente) est omise ; inversement, lorsque cette variable peut être prise en compte (eg **variable observable**), son influence sur l'endogène peut alors annihiler la présence des aberrations (il peut en aller de même en sens inverse) ;

(c) des **déviations (ou perturbations) importantes** se réalisent tout de même, quelle que soit la loi génératrice (à queues épaisses ou non) : cette situation est « rare », mais possible car la probabilité de ces **déviations**, quoique a priori faible, n'est pas nulle.

On peut ainsi distinguer deux situations a priori (cf aussi **forme d'une loi**).

(ii) **Loi à queues épaisses** (cf schéma ci-dessous). L'**expérience aléatoire** ($\mathscr{X}$, $\mathscr{B}$, P^{ξ}) considérée est l'image d'un espace probabilisé (Ω, $\mathscr{T}$, P) par une **variable aléatoire** $\xi : \Omega \mapsto \mathscr{X}$. Au lieu d'une loi à queues fines attendue, la « vraie » loi P^{ξ} de ξ, qui génère les observations, est une loi à queues épaisses, ie qui charge de façon

non négligeable des parties B $\in \mathcal{B}$ éloignées d'une **caractéristique** de **centralité** de P^{ξ} (ou d'une partie centrale C $\in \mathcal{B}$ et P^{ξ}).



Dans ce cas, les observations de ξ ne sont pas, par nature, « anormales » puisqu'elles sont engendrées par le schéma aléatoire lui-même. Par suite, tout **échantillon** $X = (X_1, ..., X_N)$ issu de ξ peut comporter des coordonnées éloignées du centre de P^{ξ}. Si les deux queues de P^{ξ} sont épaisses et si $X^{(.)} = (X^{(1)}, ..., X^{(N)})$ est l'échantillon ordonné associé à X (cf **statistique d'ordre**), il peut exister deux suites de valeurs $(X^{(1)}, ..., X^{(L)})$ et $(X^{(M)}, ..., X^{(N)})$ éloignées de la partie centrale C $\in \mathcal{B}$ de P^{ξ}.

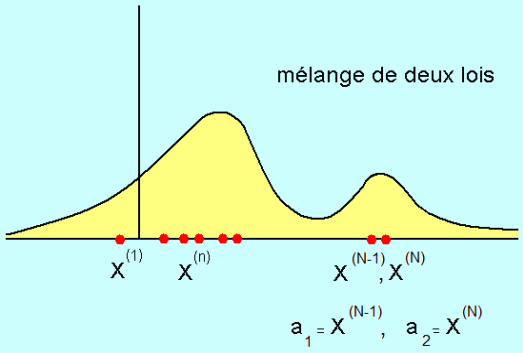
Ainsi, si $P^{\xi} = \mathcal{C}(0, 1)$ (**loi de CAUCHY** dont un **paramètre de position** est la **médiane** C $\xi = Q_{1/2}\ \xi = 0$), alors il existe a $\geq$ C ξ et b $>$ a tq, en notant $B = |a, b| \in \mathcal{B}_{\mathbf{R}}$ un intervalle quelconque de **R**, on ait $P^{\xi}(B) >> 0$. Des valeurs X_n de ξ seront observées dans B avec une **fréquence empirique** $f_{N,B}$ généralement voisine de la **proportion** $P^{\xi}(B)$.

Le type de situation précédent n'implique donc pas de traiter un problème d'aberration, sauf à remplacer la loi à queues épaisses par une loi à queues plus fines.

(iii) **Mélange de lois** (cf schéma ci-dessous). Au lieu d'observer des données conformes à une loi (« attendue ») P^{ξ}, le **statisticien** observe des données « gouvernées », en réalité, par un mélange de lois. Autrement dit, il existe $(\alpha, \beta) \in S_{n+1}$ (**simplexe** de $\mathbf{R}^{n+1}$) tq $\xi \sim M^{\xi}$, avec :

(1) $\quad M^{\xi} = \alpha P^{\xi} + \Sigma_{i=1}^{n} \beta_i P_i^{\xi}$, $\qquad$ où $\beta = (\beta_1, ..., \beta_n)$.

Loi et échantillon comportant des aberrations



Ce mélange est donc la « vraie » loi régissant les données. Par définition, $\alpha \geq 0$, $\beta_i \geq 0$ et $\alpha + \Sigma_{i=1}^{n} \beta_i = 1$. Si l'**erreur de spécification** relative à P^{ξ} n'est pas trop importante, on a aussi $\alpha >> 0$.

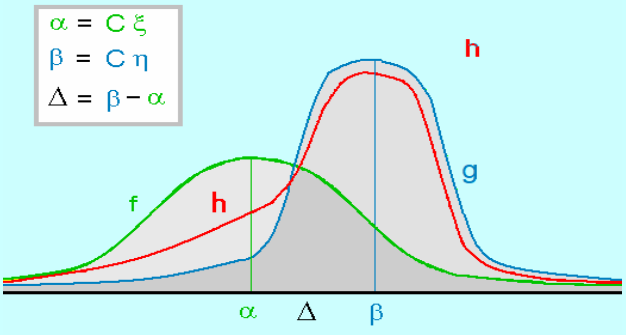
(iv) On remarque que, si le mélange M^{ξ} contient une loi à queues épaisses, eg P_1^{ξ}, le mélange est aussi à queues épaisses : ainsi, une composante P_i^{ξ} sans moments « pollue » le mélange car ce dernier sera sans moments.

(v) En pratique, on peut rencontrer les **situations statistiques** suivantes :

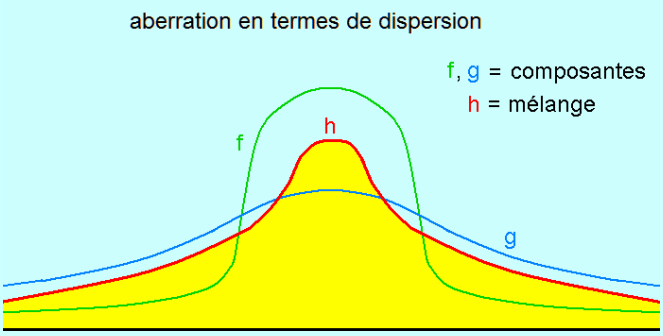
(a) **aberration en tendance centrale** (seule) : déplacements ou écarts important entre les valeurs attendues et certaines des valeurs observées. Ce déplacement doit être suffisant pour être discernable au sein du mélange (cf schéma ci-dessous) ;

aberrations en termes de centralité

lois absolument continues
mélange de deux lois de probabilité
à partir de leurs densités



(b) **aberration en dispersion** (seule) : forte différence entre la **variabilité** des valeurs attendues et celle de certaines valeurs observées. Cette situation est parfois plus complexe à détecter car le mélange ne présente pas nécessairement une forme de nature à exhiber ce type d'anomalie (cf schéma ci-dessous).



(c) **aberration par décentrage et hétérogénéité** (combinaison des deux situations précédentes).

(vi) Souvent, le mélange ne concerne que deux lois et ceci définit le **modèle à « erreur grossière »**. L'échantillon observé $Z = (Z_1, ..., Z_N)$ est alors constitué :

(a) d'une **suite iid** constituée de N_1 **copies** indépendantes de $\xi : \Omega \to \mathbf{R}^K$ dont la loi est P^ξ ;

(b) d'une suite iid constituée de N_2 copies d'une variable aléatoire $\eta : \Omega \mapsto \mathbf{R}^K$ dont la loi est P^η (avec $N_1 + N_2 = N$).

Les nombres N_1 et N_2 sont en général inconnus, ainsi que les coordonnées $X_{i1}, ..., X_{iN(1)}$ générées par P^ξ et celles X_j ($j \notin \{ i_1, ..., i_{N(1)}\}$) générées par P^η.

Ainsi, lorsque $P^\xi = \mathcal{N}_K(\mu, \Sigma)$ (**loi normale multidimensionnelle**), on peut avoir à analyser des situations tq les suivantes :

(a) **aberration en tendance centrale**. Un exemple d'aberration dûe à un **décentrage** entre moyennes est le suivant :

(2) $\quad P^\xi = \alpha \,.\, \mathcal{N}_K(\mu, \Sigma) + (1 - \alpha) \,.\, \mathcal{N}_K(\mu + \Delta\mu, \Sigma), \qquad$ avec $\alpha \in]0, 1[$.

Autrement dit, $P^\eta = \mathcal{N}_K(\mu + \Delta\mu, \Sigma)$, avec $\Delta\mu \neq 0$ (parfois $\Delta\mu >> 0$) : certaines observations sont décentrées, mais ont ici même **dispersion** que les autres ;

(b) **aberration en dispersion**. Un exemple d'aberration dûe à l'**hétérogénéité de la dispersion** (eg différence entre variances) est le suivant :

(3) $\quad P^\xi = \alpha \,.\, \mathcal{N}_K(\mu, \Sigma_1) + (1 - \alpha) \,.\, \mathcal{N}_K(\mu, \Sigma_2), \qquad$ avec $\alpha \in]0, 1[$.

Une proportion α des observations issues de P^ξ a pour **matrice de covariance** Σ_1 , et une proportion $1 - \alpha$ une matrice Σ_2 qui peut être très différente de Σ_1 .

Une même centralité des lois peut rendre délicate le traitement statistique de la différence de dispersion ;

(c) **aberration par décentrement et hétérogénéité**. Un exemple combinant les situations précédentes est le suivant :

(4) $P^{\xi} = \alpha \, . \, \mathcal{N}_K (\mu, \Sigma_1) + (1 - \alpha) \, . \, \mathcal{N}_K (\mu + \Delta \mu, \Sigma_2)$, avec $\alpha \in]0, 1[$ et $\Delta \mu \neq 0$.

Une proportion α des observations issues de P^{ξ} a pour paramètres (μ, Σ_1), et une proportion $1 - \alpha$ admet pour paramètres $(\mu + \Delta \mu, \Sigma_2)$, avec $\Sigma_2 \neq \Sigma_1$.

(vi) L'existence d'aberrations peut donc fausser la mise en oeuvre d'une **procédure statistique** (eg **estimation**, **test d'hypothèses**, **prévision**, **classification**). Elle nécessite donc une adaptation de cette procédure.

Dans un cadre général, on considère comme « vrai » (ou correctement spécifié) un **modèle image** $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$. Or les observations X ne sont pas engendrées par $\mathcal{P}^X$ mais par une famille de lois $\mathcal{Q}^X$, ie des éléments P^X de $\mathcal{P}^X$ sont des mélanges d'éléments Q^X de $\mathcal{Q}^X$.

Par suite, certaines observations (coordonnées de X) peuvent paraître « typiques » parce qu'on les pense générées par une loi P^X alors qu'elles proviennent d'une loi M^X (cf **robustesse**, **sélection de modèles**, **spécification** de modèles).

Si l'on pense que $\xi \sim P^{\xi}$ alors que $\xi \sim M^{\xi}$, avec $M^{\xi} = \alpha \, P^{\xi} + (1 - \alpha) \, Q^{\xi}$, le **rapport des vraisemblances** $\lambda = g / f$, avec $f = dP^{\xi} / d\mu$ et $g = M^{\xi} / d\mu$, est un indicateur qui intègre au numérateur l'existence d'une aberration possible pr à P^{ξ}. Le rapport des vraisemblances maximales peut alors permettre de définir un **test d'aberration** fondé sur une **statistique** de la forme $L_N = g (X) / f (X)$. Ainsi, si ξ prend de « grandes » valeurs (ie $\xi >> 0$), une **région critique** de **seuil** α sera de la forme $w = \{L_N \geq q_{1 - \alpha}\}$.

Les méthodes relatives aux mélanges de lois peuvent être adaptées (cf **dissection d'un mélange de lois**).

En pratique, les méthodes de traitement sont essentiellement des méthodes robustes (eg **régression robuste**) (cf **robustesse**), des **tests non paramétriques** de détection, qui peuvent être préalables à l'estimation d'un modèle (cf **estimateur à test préliminaire**).

L'estimation par la **méthode du maximum de vraisemblance** peut, le cas échéant, être utilisée pour des lois à extrémités épaisses (loi de CAUCHY).

Le problème de l'observation de grandeurs qui sortent du cadre d'un schéma probabiliste attendu peut donc se traiter globalement. Alternativement, on décompose souvent la méthode de résolution en deux étapes :

(a) **détection des aberrations** dans un ensemble d'observations (**test d'hypothèses**) ;

(b) puis, le cas échéant, **correction des aberrations** par des grandeurs plus vraisemblables, c'est-à-dire dont on pense qu'elles sont issues du schéma probabiliste supposé (**estimation**).

Lorsque des aberrations sont suspectées, une question d'intérêt peut porter sur la (les) population(s) qui en sont à l'origine, population(s) différente(s) de celle que l'on pense étudier (**contamination des données**).

Ce problème est à rapprocher du problème de **discrimination** mais se distingue du **problème à plusieurs échantillons** (pour lequel on connaît a priori la provenance des observations).

(vii) Le modèle de régression est un exemple dans lequel des aberrations peuvent intervenir. Ainsi, avec un **modèle de régression** non linéaire où l'échantillon (inobservable) $u = (u_1 ,..., u_N)$ est contaminé par décentrage (cf **contamination des lois**), la variable endogène η est observée selon le **vecteur aléatoire** $y = (y_1 ,..., y_N)$, qui est donc aussi contaminé en raison de la relation $y = F(b) + u$. Pour détecter les points aberrants, on examine généralement les plus fortes coordonnées (en valeur absolue) (cf **valeur extrême**) du résidu $u^{\sim} = y - F(b^{\sim})$, où le paramètre b est supposé estimé par $b^{\sim}$ selon un certaine méthode (**maximum de vraisemblance**, **régression robuste**, etc), et l'on en élimine une proportion $\alpha \in]0,1[$ déterminée.

Dans cet exemple, on peut aussi procéder par étapes, eg :

(a) soit éliminer les y_n de mêmes indices que les $u_n^{\sim}$ considérés comme aberrants, puis effectuer une seconde régression qui conduit à un estimateur $b^{\approx}$, censé être meilleur, puis à estimer les y_n contaminés par $y_n^{\approx} = F_n(b^{\approx})$;

(b) soit considérer les y_n de mêmes indices que les u_n contaminants comme des observations manquantes de η et utiliser une méthode de **régression avec lacunes**.

Une aberration peut, dans ce contexte, s'expliquer par l'**omission de variables** exogènes ou par une spécification inadaptée du modèle.

Cette situation peut d'ailleurs se présenter avec un schéma « déterministe » : cas d'une relation $\eta = f(\xi)$ vérifiée par la (N,K)-**matrice des observations** (X , y) du **couple aléatoire** (ξ , η) sans l'être par la N+1-ième observation (X_{N+1} , y_{N+1}). Dans ce cas, la prise en compte d'une K+1-ième variable exogène ξ_{K+1} associée à une extension $y_n = f^*(\xi_n , \xi_{n,K+1})$ du shéma précédent peut conduire à vérifier le « schéma étendu » $y_n = f^*(X_n , x_{n,K+1})$, non seulement pour n = 1 ,..., N, mais aussi pour n = N+1.

(viii) La notion d'aberration peut concerner aussi bien une variable numérique qu'une **variable qualitative**, mais la détection est parfois délicate. Ainsi, lorsqu'une variable (simple) κ de ce type prend ses valeurs dans l'ensemble de modalités $\mathcal{K} = \{k_1 ,..., k_M\}$ avec les probabilités correspondantes $\{q_1 ,..., q_M\}$, on peut rencontrer les situations suivantes :

(a) une probabilité q_{max} est significativement plus importante que les autres ($q_{max} >> q_m$, $\forall$ $m \neq q_{max}$), indépendamment de la modalité k_{max} associée, ne signifie pas nécessairement que k_{max} soit une modalité anormale ;

(b) alternativement, une modalité k_a donnée peut paraître « atypique » (comparée aux autres), mais sa probabilité q_a n'est pas nécessairement négligeable.

Autrement dit, les modalités aussi bien que leurs probabilités n'informent pas, en général, sur le caractère aberrant ou non de l'une ou plusieurs d'entre elles.