DÉCOMPOSITION DE LA VARIANCE (A3, C5, D1, F3, L, M)
(08 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

La **variance** (théorique) d'une **variable aléatoire**, ou la **variance empirique** d'un **échantillon** (cf **dispersion**), peut s'exprimer de plusieurs façons en fonction de « composantes » (internes ou externes) significatives.

(i) Deux formules, l'une théorique et l'autre empirique, sont directement issues de la **formule de KOENIG-HUYGENS**.

(ii) La variance d'une va peut aussi se décomposer par **conditionnement**. Ainsi, soit $(\xi, \eta) : \Omega \mapsto \mathbf{R}^2$ un **couple aléatoire** de carré intégrable, défini sur un **espace probabilisé** $(\Omega, \mathcal{T}, P)$. On a alors la **formule de décomposition de la variance** suivante :

(1) $\quad V\,\eta = V_\xi\, E\,(\eta / \xi) + E_\xi\, V\,(\eta / \xi),$

où $V\,\eta$ désigne la variance propre de η, $E\,\eta / \xi$ l'espérance de η conditionnelle à ξ, $V\,(\eta / \xi)$ la variance de η conditionnelle à ξ, E_ξ est l'espérance de la va $V\,(\eta / \xi)$ (comme fonction de la **va** ξ) et V_ξ est la variance de la va $V\,(\eta / \xi)$ (comme fonction de ξ).

La variance de η se décompose donc en somme d'une variance de moyennes conditionnelles et d'une moyenne de variances conditionnelles (la démonstration est basée sur le **lemme de BLACKWELL**).

(iii) La formule s'étend directement à un couple de **vecteurs aléatoires** $(\xi, \eta) : \Omega \mapsto \mathbf{R}^K \times \mathbf{R}^G$.

(iv) Si la formule (1) est calculée en remplaçant formellement la loi $P^{(\xi, \eta)}$ de (ξ, η) par la **loi empirique** associée à un N-**échantillon** $Z = (Z_1, ..., Z_N)$ issu de ce couple, avec $Z_n = (X_n, Y_n)$, on obtient une formule empirique analogue (cf **statistique naturelle**).

(v) Si $(\xi, \eta, \zeta) : \Omega \mapsto \mathbf{R}^3$ est un triplet de **vars**, on montre, de même, que :

(2) $\quad V\,\zeta = V_\eta\, E_{\xi\,/\,\eta}\,(E\,\zeta / (\xi, \eta)) + E_\eta\, V_{\xi\,/\,\eta}\,(E\,\zeta / (\xi, \eta)) + E_\eta\, E_{\xi\,/\,\eta}\,(V\,\zeta / (\xi, \eta))$

(avec des notations analogues aux précédentes).

(vi) Une écriture symbolique permet d'alléger les notations. Ainsi, (1) s'écrit :

(1)' $\quad V = V_2\, E_1 + E_2\, V_1\ ,$

(2) s'écrit :

(2)' $\quad V = V_3\, E_2\, E_1 + E_3\, V_2\, E_1 + E_3\, E_2\, V_1\ .$

La formule symbolique générale s'écrit, avec k degrés de conditionnement :

(3) $\quad V = \Sigma_{i=1}^{k}\, E_k \ldots E_{i+1}\, V_i\, E_{i-1} \ldots E_1\ ,$

avec les conventions $E_{k+1} = V_k$ et $E_0 = V_1$. On peut ainsi l'interpréter comme un opérateur de conditionnements successifs.

Des propriétés tq (1) ou (2) sont utilisées en **théorie des sondages** : en effet, dans un **sondage à plusieurs degrés**, les conditionnements de chaque degré sont effectués pr aux **plans de sondage** relatifs aux degrés supérieurs.

(vii) Une autre formule de décomposition de la variance (théorique ou empirique) consiste en la **décomposition spectrale** d'une **matrice de covariance** (ou d'un **opérateur de covariance**).

(viii) En **Statistique descriptive** ou en **théorie des sondages**, on emploie souvent une **formule de décomposition de la variance** d'un autre type, **associée à une partition** donnée d'une **population** finie. Soit Ω une population finie (Card Ω = M), $\Pi_\Omega = \{\Omega_1 ,..., \Omega_H\}$ une **partition** de Ω et $\eta : \Omega \mapsto \mathbf{R}$ un **caractère** donné (**variable statistique** ou **va**), mesuré sur les individus $\omega_m \in \Omega$, $\forall$ m $\in \{1 ,..., M\}$. On pose :

$Y = (Y_1 ,..., Y_M)'$, avec $Y_m = \eta(\omega_m)$, $\forall$ m $\in N_M^*$;

$\bar{Y}_M = M^{-1} \Sigma_{m=1}^{M} Y_m = e_M' Y / M$ (**moyenne arithmétique** d'ensemble) ;

$S = Y' P_M Y$ (**variance** d'ensemble) ;

(4) $\text{Card } E_h = M_h \neq 0$, $\forall$ m $\in N_M^*$;

$Y^h = (Y_1^h ,..., Y_{M(h)}^h)'$, avec $Y_m^h = \eta(\omega_m) \Leftrightarrow \omega_m \in \Omega_h$, $\forall$ m $\in \{1 ,..., M_h\}$, $\forall$ h ;

$\bar{Y}^h = e_{M(h)}' Y^h / M_h$ (moyenne de la classe (**strate**) h) ;

$S^h = (Y^h)' P_{M(h)} Y^h / M_h$ (variance de la classe (**strate**)).

en notant aussi M(h) pour M_h .

Alors, la **formule de décomposition de la variance totale** sécrit :

(5) $S = \Sigma_{h=1}^{H} (M_h / M) (\bar{Y}^h - \bar{Y}_M)^2 + \Sigma_{h=1}^{H} (M_h / M) S^h$.

L'interprétation est analogue à celle de la formule (1).

Il existe une formule semblable pour la variance « empirique » calculées sur un N-**échantillon** $A = \{a_1 ,..., a_N\}$ extrait de Ω.

Dans l'étude des **plans d'expérience** (où les strates précédentes sont des **blocs**), on rencontre encore une formule analogue à (5).

Ce dernier type de décomposition est souvent employé conjointement avec le **théorème de COCHRAN** et les propriétés de **convolution** de la **loi du chi-deux** (cf **équation d'analyse de la variance**, **partition du chi-deux**).

(ix) Enfin, le **modèle à variance composée** et le modèle de **décomposition de la variance** sont deux exemples d'étude statistique d'une variance (ou d'une dispersion).