VARIANCE (C5, F3)
(08 / 06 / 2020, © Monfort, Dicostat2005, 2005-2020)

Alors que la **moyenne** d'une **variable aléatoire** indique une « position » (cf **centralité**, **paramètre de position**), la **« variance »** (R.A. FISHER) renseigne sur sa « **dispersion** » relativement à cette position. Ces deux premiers **moments** sont très utilisés, ensemble ou séparément (cf aussi **variable centrée**, **variable réduite**, **variable normée**).

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé**, $(\mathbf{R}, \mathcal{B}_\mathbf{R})$ la droite réelle probabilisable et $\xi : \Omega \mapsto \mathbf{R}$ une **vars** de **loi** P^ξ. On suppose ξ de carré intégrable, ie $\xi \in \mathcal{L}_\mathbf{R}^2 (\Omega, \mathcal{T}, P)$, et l'on note $E\,\xi$ son **espérance mathématique**.

On appelle alors **variance théorique**, ou **dispersion théorique**, de ξ le nombre réel positif (qui existe donc) suivant :

(1) $\quad V\,\xi$ ou $\sigma^2 = E\,(\xi - E\,\xi)^2 = \int (\xi - E\,\xi)^2\, dP = \int (x - E\,\xi)^2\, dP^\xi\,(x)$.

La variance est donc le moment théorique (d'ordre 2) centré pr à la **moyenne** (théorique) $E\,\xi$.

Une définition alternative de la variance est la suivante :

(3) $\quad V\,\xi$ ou $\sigma^2 = (1/2)\, E\,(\xi' - \xi'')^2 = (1/2) \int (\xi' - \xi'')^2\, dP$,

ou ξ' et ξ'' sont deux va indépendantes et de même loi P^ξ que ξ (**couple aléatoire** iid).

On appelle alors **écart-type théorique** la racine carrée positive de la variance (théorique) : $\sigma = (V\,\xi)^{1/2}$.

(ii) La variance vérifie diverses propriétés, dont les plus élémentaires :

(a) $V\,\xi \geq 0$;

(b) $V\,\xi = 0 \Rightarrow$ il existe une constante $c \in \mathbf{R}$ tq $\xi = c$ (P-p.s.).

(c) **formule d'inertie de S. KOËNIG** (cf **formule de KOENIG-HUYGENS**) :

(2) $\quad V\,\xi = E\,\xi^2 - (E\,\xi)^2$.

(iii) Soit $X = (X_1, ..., X_N)'$ un **échantillon iid** généré par ξ (ie indépendant et équidistribué selon P^ξ) et soit P_N la **loi empirique** associée à X.

On appelle **variance empirique**, ou **dispersion empirique** de ξ (ou de X) le nombre aléatoire :

(4) $V_N\,\xi$ ou S_N^2 ou $V_N\,X = N^{-1} . \Sigma_{n=1}^{N} (X_n - \bar{X}_N)^2$,

obtenu en remplaçant P^{ξ} par P_N dans le calcul de la variance théorique (1) (cf **statistique naturelle**).

La variance empirique est encore notée $v_N\,\xi$ (ou $v_N\,X$) ou simplement $v\,\xi$ (ou $v\,X$). On peut l'exprimer sous l'une des formes suivantes :

(3)' $S_N^2 = N^{-1}\,||X' - \bar{X}_N\,e_N||^2 = (e_N'\,e_N)^{-1}\,X'\,P\,X$,

où X est le **vecteur aléatoire** (en colonne) représentant l'échantillon et $P \in M_N(\mathbf{R})$ la **matrice de centrage par rapport à la moyenne** $\bar{X}_N$, ie $P = I_N - J_N$, avec $J_N = (e_N'\,e_N)^{-1}\,(e_N\,e_N')$.

Par suite :

(5) $S_N^2 = N^{-1}\,(||X||^2 - N^{-1}\,X'\,e_N\,e_N'\,X)$.

On appelle **écart-type empirique** la racine carrée positive de la variance empirique, ie S_N .

(iv) La variance empirique S_N^2 est un **estimateur** biaisé de σ^2 (cf **estimateur sans biais**) :

(6) $E\,S_N^2 = N^{-1}\,(N-1)\,\sigma^2 = N^{-1}\,(N-1)\,V\,\xi$.

On peut cependant en déduire un estimateur sans biais, appelé **variance corrigée**, selon :

(7) $(S_N^2)^* = (N-1)^{-1}\,N\,S_N^2 = (N-1)^{-1}\,\Sigma_{n=1}^{N} (X_n - \bar{X}_N)^2$.

Dans le cas où $\xi \sim \mathcal{N}_1(\mu, \sigma^2)$ (**loi normale**), on montre que :

(8) $N\,(S_N^2 / \sigma^2) \sim \mathcal{X}_{N-1}^2$ (**loi du chi-deux** à N - 1 **degrés de liberté**).

Ceci permet d'effectuer des **tests d'hypothèses** relatifs à une variance.

(v) Dans le problème à deux échantillons (cf **problème à plusieurs échantillons**), on considère un **échantillon iid** $X^i = (X_{i1}, ..., X_{iN(i)})$ dont la **variable parente** est une **vars** ξ_i (i = 1, 2) (en notant n(i) pour désigner n_i et N(i) pour désigner N_i).

Par suite, si ξ_1 est indépendant de ξ_2 et si $\xi_i \sim \mathcal{N}_1(\mu_i, \sigma_i^2)$ (i = 1, 2) (variables gaussiennes), alors, la **statistique** :

(9) $F_{N(1)N(2)} = (S_{N,1}^2)^* / (S_{N,2}^2)^*$

vérifie :

(10) $F_{N(1)N(2)} / \rho_{12}^2 \sim \mathcal{F}(N_1 - 1 , N_2 - 1)$

(**loi de FISHER-SNEDECOR** à $N_1 - 1$ et $N_2 - 1$ degrés de liberté), où l'on note $(S_{N,i}^2)^* = (N_i-1)^{-1} \Sigma_{n(i)=1}^{N(i)} (X_{in(i)} - \bar{X}_{N,i})^2$ (i = 1, 2) la variance empirique corrigée associée à X^i, $\bar{X}_{N,i}$ sa moyenne empirique (i = 1, 2) et $\rho_{12}^2 = \sigma_1^2 / \sigma_2^2$ le **rapport des variances** théoriques.

Ce résultat sert à effectuer des **tests** relatifs à ρ_{12}^2.

Si les variables ne sont pas gaussiennes, les propriétés précédentes peuvent être altérées (absence de **robustesse** en présence d'écarts pr à la **normalité**).

La notion de variance se généralise, dans le cas multidimensionnel, en celle de **dispersion**.