**VARIABLE ALÉATOIRE (C1)**
(18 / 10 / 2021, © Monfort, Dicostat2005, 2005-2021)

[**Note**. On donne souvent à une **variable aléatoire** le même qualificatif que le nom de sa **loi de probabilité** : variable gaussienne et **loi gaussienne**, variable de POISSON - ou variable poissonnienne - et **loi de POISSON**, etc]

Notion fondamentale du **calcul des probabilités**, la notion de **variable aléatoire** correspond à la notion mathématique (théorie de la mesure) d'**application mesurable**.

En **Statistique**, on utilise plutôt le concept « technique » de **statistique**, en association avec les notions de **modèle statistique** et de **problème statistique**.

(i) Soit $(\Omega, \mathcal{T})$ et $(\mathcal{X}, \mathcal{B})$ deux **espaces mesurables** (ou **espaces probabilisables**) et $\xi : \Omega \mapsto \mathcal{X}$ une **application** donnée.

On dit que $\xi$ est une **variable aléatoire**, ou un **aléas**, ou un **élément aléatoire**, défini sur $\Omega$ et à valeurs dans $\mathcal{X}$, ssi $\xi$ est une application $(\mathcal{T},\mathcal{B})$-mesurable, ie ssi $\xi$ est compatible avec les **structures** mesurables $(\Omega, \mathcal{T})$ et $(\mathcal{X}, \mathcal{B})$ resp associées à $\Omega$ et à $\mathcal{X}$. Autrement dit, par définition :

(1) $\quad \xi^{-1}(B) \in \mathcal{T}, \quad \forall B \in \mathcal{B}, \quad$ ou encore $\quad \xi^{-1}(\mathcal{B}) \subset \mathcal{T},$

ie l'image inverse de la **tribu** $\mathcal{B}$ par $\xi$ est contenue dans la tribu $\mathcal{T}$.

Une variable aléatoire (ou **va**) se note à l'aide de symboles variés : eg des lettres grecques $\xi$, $\eta$, $\zeta$ ou $\varepsilon$ (etc), ou des lettres latines majuscules X, Y, Z ou U, etc :

(a) les lettres grecques peuvent servir à représenter des va « théoriques » (eg des **inobservables**) ;

(b) les lettres latines représentent des va « empiriques », ie des variables constituant des **observations** des va précédentes.

Ainsi, $\xi$ représentera une **vars** et $X = (X_1, ..., X_N)$ un échantillon constitué de **copies** (ou « répliques ») analogues à la « **variable parente** » $\xi$, eg des copies indépendantes entre elles et équidistribuées comme $\xi$ (cf **échantillon iid**).

Plus généralement :

(a) un **échantillon aléatoire** quelconque $X = (X_1, ..., X_N)$ peut aussi être considéré comme une va à valeurs dans un espace produit $\mathcal{X} = \Pi_{n=1}^{N} \mathcal{X}_n$, ou dans un espace puissance $\mathcal{X} = \mathcal{X}_0^N$, ou encore dans un **espace vectoriel** (réel) $\mathcal{X}$ ;

(b) de même, un **processus stochastique** peut être considéré comme une variable aléatoire $X = (X_t)_{t \in T}$ à valeurs dans un espace produit (en général infini) $\mathcal{X} = \Pi_{t \in T} \mathcal{X}_t$ , ou dans un espace puissance $\mathcal{X} = \mathcal{X}_0^T$ .

Si $\xi$ est une va, on note souvent $x = \xi(\omega)$ l'image d'un élément $\omega \in \Omega$ par $\xi$. On note encore $x = X(\omega)$ l'image d'un élément $\omega \in \Omega$ par X : cette image (ou « valeur ») est appelée **observation** lorsqu'elle peut être « appréhendable » sur des **unités statistiques** (cf **observabilité**, **variable observable**). On l'appelle encore « mesure », quelle que soit la nature de la variable considérée.

(ii) Une va est généralement qualifiée en fonction de la nature (mathématique) de l'**ensemble** de ses valeurs. Ainsi, on distingue trois principaux types de variables :

(a) $\xi : \Omega \mapsto \mathcal{X}$ est une **variable aléatoire quantitative**, ou **variable aléatoire numérique**, (scalaire) ssi l'ensemble $\mathcal{X}$ de ses valeurs est un ensemble numérique (cf **variable quantitative**). Parmi les variables numériques :

$(a)_1$ $\xi$ est appelée **variable discrète** ssi $\mathcal{X}$ est un **ensemble discret**, ie un ensemble ayant au plus la **puissance du dénombrable** (ie soit fini, soit en bijection avec l'ensemble **N** ou l'une de ses parties (cf **loi discrète**). Une variable de cette nature peut encore être qualifiée de **variable « simple »**, ou de **variable** « **scalaire** » ou de **variable** « **à une dimension** ».

Ainsi, on dit que $\xi$ est :

∗ une **variable entière non négative** lorsque $\mathcal{X} = \mathbf{N}$ ; une variable entière non négative (finie ou bornée) lorsque $\mathcal{X} = \{0, 1, ..., n\} = N_n$ ;

∗ une **variable entière positive** (ou **variable** « **naturelle** ») lorsque $\mathcal{X} = \mathbf{N}^* = \mathbf{N} \setminus \{0\}$ ; une variable entière positive (finie ou bornée) lorsque $\mathcal{X} = \{1, ..., n\} = N_n^*$ ;

∗ une **variable entière signée** lorsque $\mathcal{X} = \mathbf{Z}$ ; une variable entière signée (finie ou bornée) lorsque $\mathcal{X} = \{-m, ... , 0 , ... +n\} = Z_{mn}$ ;

∗ une **variable rationnelle signée** lorsque $\mathcal{X} = \mathbf{Q} = \mathbf{Z} / \mathbf{N}^*$.

Par extension, on peut aussi considérer comme discrète une variable « continue » dont la **loi de probabilité** ne « charge » qu'un ensemble discret $\mathcal{X}$ du type ci-dessus : cette loi est parfois appelée **« distribution de masses »**.

$(a)_2$ $\xi$ est appelée **variable continue** lorsque l'ensemble $\mathcal{X}$ de ses valeurs possède la puissance du continu. Une variable de cette nature est parfois appelée **variable « scalaire »** ou **variable « à une dimension »**.

Ainsi, on dit que $\xi$ est :

∗ une **variable réelle signée** lorsque $\mathscr{X}$ = **R** ;

∗ une **variable réelle non négative** lorsque $\mathscr{X}$ = $\mathbf{R}_+$ = {x $\in$ **R** : x $\geq$ 0} ; une **variable réelle positive** lorsque $\mathscr{X}$ = $\mathbf{R}_+^*$ = {x $\in$ **R** : x $>$ 0} = {x $\in$ $\mathbf{R}_+$ \ {0}} ; il en va de façon analogue pour les valeurs dans $\mathbf{R}_-$ ou $\mathbf{R}_-^*$ ;

∗ une **variable complexe** lorsque $\mathscr{X}$ = **C**.

Lorsque $\mathscr{X}$ = **R**, on parle parfois d'**aléa numérique**. En pratique, le plus souvent (cas scalaire), $\mathscr{X}$ = **D** (ensemble des **nombres décimaux**) : en effet, la « représentation machine » d'un nombre mémorisé dans un ordinateur (mémoire vive ou mémoire morte) ne peut, pour des raisons matérielles, qu'être tronquée, même si sa précision est très élevée ;

(b) $\kappa : \Omega \mapsto \mathscr{K}$ est une **variable aléatoire qualitative**, ou une **variable aléatoire attribut**, ssi l'ensemble $\mathscr{K}$ de ses valeurs n'est doté d'aucune propriété en termes de calcul algébrique (cf **variable qualitative**). Autrement dit, $\mathscr{K}$ est un **ensemble non numérique**, ie un ensemble dont les éléments ne peuvent faire l'objet de calculs algébriques. Dans certains cas, les éléments de $\mathscr{K}$ peuvent avoir une « apparence numérique » (eg identifiant, numéro d'ordre, etc), mais les calculs entre ces éléments sont insensés.

Une variable de cette nature peut aussi être qualifiée de « **variable simple** », ou de « **variable scalaire** », ou encore de « **variable à une dimension** » .

On distingue usuellement entre :

$(b)_1$ **variable nominale**, pour laquelle $\mathscr{K}$ ne possède pas de structure particulière ;

$(b)_2$ **variable ordinale**, pour laquelle $\mathscr{K}$ est un ensemble ordonné (cf **relation d'ordre**), qui peut lui-même résulter d'un ensemble numérique ;

(c) **variable morphologique** ou **variable objectale**, $\xi : \Omega \mapsto \mathscr{K}$ tq l'ensemble $\mathscr{K}$ de ses valeurs est constitué de « formes » ou d'« objets » particuliers (cf **forme**, **reconnaissance des formes**), eg :

$(c)_1$ $\mathscr{K}$ est un ensemble de figures ou de formes géométriques : coniques, quadriques, ou, plus généralement, **variétés différentielles**, **graphes**

(**arborescence** ou **réseau**), etc. Cet **ensemble** correspond aux valeurs prises par une **variable aléatoire géométrique** ;

$(c)_2$ $\mathcal{K}$ est un ensemble (ou « espace ») de **fonctions numériques**. Cet espace correspond à une **variable aléatoire fonctionnelle**.

Dans cette dernière situation, les formes $k \in \mathcal{K}$ (ensemble considéré dépendent souvent d'une variable (vectorielle réelle) $\theta$, ie $\theta \mapsto k = \phi(\theta)$, avec $\theta \in \Theta \subset \mathbf{R}^Q$ (où $Q \in \mathbf{N}^*$). La **loi multidimensionnelle** (**loi jointe**) de $\theta$ détermine alors le caractère aléatoire de la forme $\phi$ : l'analyse de cette situation se ramène à celle de la loi du **vecteur aléatoire** $\theta$, donc de la loi d'une variable quantitative vectorielle (cf infra).

Si les modalités constituant $\mathcal{K}$ sont déterminées par une « **morphologie** », alors $\mathcal{K}$ est un ensemble de formes que l'on peut distinguer des deux types précédents (sauf eg lorsque $\mathcal{K}$ est lui-même une classification déduite d'un ensemble de formes).

On peut ainsi considérer que la notion de variable morphologique est une notion intermédiaire entre celle de variable qualitative (puisqu'on ne peut effectuer de calculs entre formes) et celle de variable numérique (puisque les formes sont souvent définies ou calculables à partir d'équations numériques tq $k = \phi(\theta)$ ci-dessus).

(iii) On appelle **suite aléatoire**, ou parfois **« liste » aléatoire**, (finie) une va $\zeta$ dont l'ensemble $\mathcal{Z}$ des « valeurs » est un **ensemble produit (fini)** de la forme $\mathcal{Z} = \Pi_{i=1}^n \mathcal{Z}_i$ . Chaque composante $\mathcal{Z}_i$ de ce produit est muni d'une **structure** mesurable, éventuellement adaptée à sa nature, ou à sa structure, particulière : $\mathcal{Z}$ peut ainsi constituer un « mélange » d'ensembles numériques ou non, géométriques ou non, unidimensionnels ou non, etc.

On appelle **vecteur aléatoire**, ou **variable aléatoire vectorielle**, une va $\xi : \Omega \mapsto \mathcal{X}$ tq $\mathcal{X}$ est un **espace vectoriel** (réel) muni d'une structure mesurable (resp **espace vectoriel mesurable**). On dit notamment que $\xi$ est un **élément aléatoire** ssi $\mathcal{X}$ est un **espace de BANACH** (généralement muni de sa **tribu borélienne**).

Ainsi, lorsque $\mathcal{X}$ est un espace vectoriel sur un corps **K** (en pratique **K** = **R** ou **K** = **C**) et qu'il est muni d'une base canonique $(e_i)_{i = 1,\dots,n}$ , la variable $\xi$ se décompose selon $\xi = \Sigma_{i=1}^n \xi_i e_i$ , où les $\xi_i$ sont des va à valeurs dans **K**.

(a) si $\mathcal{X} = \mathbf{R}^n$, on dit que $\xi$ est un **vecteur aléatoire réel**, ou une **variable aléatoire vectorielle réelle** ;

(b) si $\mathcal{X} = \mathbf{C}^n$, on dit que $\xi$ est un **vecteur aléatoire complexe**, ou une **variable aléatoire vectorielle complexe** ;

En pratique, on identifie **C** à $\mathbf{R}^2$, muni de sa tribu borélienne $\mathcal{B}(\mathbf{R}^2)$.

En particulier :

(a) si $\mathcal{X} = \mathbf{K}$, on dit que $\xi$ est une **variable aléatoire réelle scalaire** (vars) si **K** = **R** ou une **variable aléatoire complexe scalaire** (vacs) si **K** = **C** ;

(b) si $\xi$ est une **variable aléatoire matricielle**, ou **matrice aléatoire**, ssi l'ensemble de ses valeurs est $\mathcal{X} = M_{mn}(\mathbf{K})$ (espace vectoriel des matrices de format (m,n), où **K** est un corps donné). Cette notion est à distinguer de celle de **matrice stochastique**.

(iv) Le concept de va est donc très générale, et il en est de même de celui de statistique (qui en résulte).

Ainsi, une va peut prendre ses valeurs dans un ensemble $\mathcal{X}$ dont les éléments sont de nature très variée, eg :

(a) **histogrammes** ;

(b) **matrices** ou **tableaux statistiques** ;

(c) figures géométriques : graphiques, diagrammes, cercles, triangles, etc (cf **probabilité géométrique**). On peut aussi considérer :

$(c)_1$ des **graphes** : on parle alors de **graphes aléatoires** ;

$(c)_2$ des **réseaux** : on parle alors de **réseaux aléatoires** ;

(v) Lorsque $(\Omega, \mathcal{T})$ est muni d'une **mesure de probabilité** définie sur $\mathcal{T}$, l'**espace probabilisé** $(\Omega, \mathcal{T}, P)$ devient, par $\xi$, un espace probabilisé $(\mathcal{X}, \mathcal{B}, P^\xi)$, ou $P^\xi$ est, par définition, la **loi de probabilité** de $\xi$ (ie l'image de P par $\xi$) (cf aussi **mesure image**). Dans ce cas, les qualificatifs précédents sont utilisés en considérant le **support** Supp $P^\xi$ de la loi de $\xi$, et non pas l'ensemble $\xi(\Omega) \subset \mathcal{X}$ des valeurs possibles de $\xi$, ie son image Im $\xi$ (image de $\Omega$ par $\xi$).

(vi) Par ailleurs :

(a) on suppose souvent que $(\mathcal{X}, \mathcal{B}) = (\Omega, \mathcal{T})$ et $\xi = id_\Omega$ (identification de l'**espace fondamental** de départ et de l'**espace d'observation** image) ;

(b) la notion de variable aléatoire ne dépend pas du fait que l'**espace probabilisable** $(\Omega, \mathcal{T})$ est muni d'une (mesure de) probabilité P. Cependant, on considère parfois l'espace probabilisé $(\Omega, \mathcal{T}, P)$ et l'ensemble $\mathcal{A}(\Omega, \mathcal{T})$ des

applications mesurables (va au sens précédent) pour définir la **relation d'équivalence** suivante sur $\mathcal{A}$ $(\Omega, \mathcal{T})$ :

(2) $\xi \sim \eta \Leftrightarrow \xi = \eta$ (P - p.s.).

On appelle alors **variable aléatoire** tout élément de l'**espace quotient** A $(\Omega, \mathcal{T}) = \mathcal{A}$ $(\Omega, \mathcal{T}) / \sim$, constitué des classes de va (au sens précédent) qui sont P-presque sûrement égales entre elles. Par suite, les concepts usuels (**espérance mathématique**, **moment**, **mode**, **quantile**, etc) sont définis à partir d'éléments de A $(\Omega, \mathcal{T})$ ;

(c) le concept de variable aléatoire généralise celui d'**événement aléatoire**. En effet, la variable aléatoire $\mathbf{1}_A$ (**variable indicatrice** de l'événement $A \in \Omega$) est une **vars** particulière, puisqu'elle prend ses valeurs dans $[0, 1] \subset \mathbf{R}$ (cf aussi **fonction indicatrice**) ;

(d) une **variable aléatoire** peut être **observable** (ie « visible ») (cf **observabilité**) ou **non observable** (ie **inobservable** ou « invisible »). Ainsi, on peut être conduit à distinguer (eg en psychologie) entre « **variables manifestes** » et « **variables latentes** ».

En effet, il arrive souvent qu'une théorie (ou un modèle), dans sa **forme initiale**, relie, sous des formes tq :

(3) $\omega_g = \rho_g (\iota_h)$ $(g = 1, ..., G)$ $(h = 1, ..., H)$

une liste $\omega_g$ de variables observables à une liste $\iota_h$ de variables inobservables, en sorte que (si l'élimination des $\iota_h$ est possible), dans la **forme finale** de cette théorie (ou de ce modèle) seules les variables observables $\omega_g$ apparaissent (ie soient manifestes) :

(4) $\sigma (\omega_1, ..., \omega_G) = 0.$

Ce procédé permet, en théorie, de réaliser l'**inférence statistique** recherchée, mais comporte des risques. Ainsi, en supposant que la « vraie » relation (inconnue) entre observables est $\omega_2 = \delta / \omega_1$ alors que la **spécification** initiale s'écrit $\omega_1 = \alpha . i_1 + \beta$ et $\omega_2 = \gamma . i_1^2$, l'élimination de $i_1$ conduit à une relation $\omega_2 = \gamma . ((\omega_1 - \beta) / \alpha)^2 + \beta$ différente de la vraie (l'une est hyperbolique, l'autre quadratique, en $\omega_1$).

(vii) On donne souvent à une variable aléatoire le **nom** de sa loi de probabilité : ainsi, on parle de **variable uniforme** (dont la loi est uniforme, qu'elle soit discrète ou continue), de variable binômiale (dont la loi est la **loi binômiale**), de **variable de POISSON** (dont la loi est la **loi de POISSON**), de **variable normale** ou de **variable gaussienne** (dont la loi est la **loi normale** ou **loi gaussienne**), etc.

(viii) Les va utilisées sont souvent des transformées de variables « initiales » : leurs « valeurs » (éléments de $\mathcal{X}$) sont donc obtenues par **changement de variable**

**aléatoire**. De même, l'observation des va sur des **unités statistiques** fait souvent, ensuite, l'objet de **transformation des données** initiales (celles de l'**échantillon**) : « tabulation » (ie **agrégation**), groupement de données en classes (cf **groupement de classes**), « représentations » graphiques, etc (cf aussi **disposition**).

(ix) La notion de variable aléatoire est généralement associée à celle d'**espace probabilisable** (ou d'espace probabilisé). Elle se généralise en celle de **statistique**, laquelle est associée à la notion de **modèle statistique**, ou de **représentation statistique**, ie de **famille** d'espaces probabilisés.