

CARACTÈRE (C1, C3)

(06 / 06 / 2019)

En **Statistique**, un **ensemble** (ou une **population**) de référence est généralement décrit(e) à l'aide de diverses **variables** (ou « descripteurs ») observées sur chaque élément (**unité statistique** ou « individu ») de cet ensemble. Il en est de même de la description d'un **phénomène** relevant d'un **domaine de connaissance** donné : eg luminosité d'un corp céleste, perméabilité d'une membrane de bactérie dans une culture, revenu d'une personne physique, etc.

Un **caractère** peut désigner une **variable statistique**, qu'elle soit numérique ou non.

(i) Qu'il soit numérique ou qualitatif, le concept sous-jacent à un caractère peut être :

(a) soit le concept mathématique (ie probabiliste) de **variable aléatoire** ;

(b) soit le concept statistique de **statistique**.

Soit Ω un **ensemble** et ω un élément de Ω . On appelle **caractère**, ou parfois **caractéristique**, sur Ω toute **variable** ξ , qui peut être de nature quantitative (cf **variable quantitative**), qualitative (cf **variable qualitative**), ou encore morphologique (cf **variable morphologique** (variable décrivant des **formes**)). Une telle variable est donc définie pour chaque $\omega \in \Omega$: si $\mathcal{X}$ dénote l'ensemble des valeurs possibles de ξ , l'application $\xi : \Omega \mapsto \mathcal{X}$ elle-même est parfois aussi appelée **caractère**.

En général, on suppose que ξ est une **application mesurable** (ce qui suppose définies une **tribu de parties** $\mathcal{T}$ sur Ω et une tribu $\mathcal{B}$ sur $\mathcal{X}$) : un caractère n'est alors autre chose qu'une **va** ou une **statistique**, et $(\mathcal{X}, \mathcal{B})$ s'interprète comme un **espace d'observation**.

(ii) Parmi les **caractères quantitatifs**, on distingue (cf **variable qualitative**) :

(a) les **caractères « continus »** (ou assimilés comme tels) : $\mathcal{X} \in \mathcal{T}(\mathbf{R})$ (intervalles de $\mathbf{R}$) (eg la distance κ d'une étoile du système solaire au soleil à une date donnée) (cf **variable continue**) ;

(b) les **caractères discontinus**, ou **caractères discrets** (ou assimilés comme tels) : $\mathcal{X}$ est en bijection avec $\mathbf{N}$ (ou avec une partie finie de $\mathbf{N}$) (eg le nombre ξ de bactéries d'une culture en fin d'expérimentation) (cf **variable discrète**).

En pratique, la distinction entre caractère discret et caractère continu est souvent conventionnelle. Ainsi, un caractère supposé prendre des valeurs continues (ie tq $\mathcal{X} = \mathbf{R}$) mais observé à travers un instrument de mesure dont la précision est limitée ne sera réellement observé que dans l'ensemble $\mathbf{Q}$ (nombres rationnels), ou seulement

dans une partie stricte de $\mathbf{Q}$ (eg l'ensemble $\mathbf{D}$ des nombres décimaux). Un caractère continu peut être « discrétisé » (cf **discrétisation**), et un caractère (numérique) discret peut être « interpolé » (cf **interpolation**).

(iii) Lorsque $\mathcal{X}$ est un ensemble de type qualitatif ou descriptif (et en général fini) $\mathcal{K} = \{k_1, \dots, k_M\}$, $\kappa : \Omega \mapsto \mathcal{K}$ est une variable qualitative, et chacune de ses « valeurs » k_m ($k \in N_M^*$) est appelée **modalité** du caractère κ . Cette application associe donc à chaque $\omega \in \Omega$ une caractéristique $k = \kappa(\omega) \in \mathcal{K}$.

Un **caractère qualitatif**, ou **attribut**, κ peut être un **caractère ordonné** (eg une variable définie à partir d'une partition de $\mathcal{X} = \mathbf{R}$) (cf **échelle ordinale**) ou un **caractère non ordonné** (cf **variable qualitative**).

Un exemple de **caractère non ordonné** est celui où $\mathcal{K}$ représente une **nomenclature**, ie une liste comportant M (nombre généralement fini) de **modalités** (ou **descriptions**), attribuables à des unités statistiques : on impose aux modalités k_m ($m = 1, \dots, M$) de cette liste des contraintes de non compatibilité et d'exhaustivité.

Une notion de **nomenclature « floue »** est aussi concevable (cf **partie floue**, **théorie des parties floues**) : ainsi, l'ensemble des activités économiques peut être décrite sous forme de filières de production non disjointes, certaines activités pouvant appartenir à des filières distinctes (activités communes).

(iv) Les modalités k_m d'un caractère numérique κ peuvent être définies par découpage de $\mathcal{X} = \mathbf{R}$ selon des « **tranches de valeurs** » $[a_{m-1}, a_m[$ (avec $m \in \{2, \dots, M\}$). On remplace alors $[a_{m-1}, a_m[$ par une valeur intermédiaire $a_{m-1} + \gamma (a_m - a_{m-1})$ (où $\gamma \in [0, 1]$).

Ce procédé, appelé **discrétisation par tranches** du caractère κ , est aussi employé lorsque $\mathcal{X} = \mathbf{R}^K$, la **partition** de chaque espace facteur $\mathbf{R}$ n'étant pas, en général, la même (du moins lorsqu'il s'agit d'une partition produit) (cf aussi **discrétisation**).

Ceci est à distinguer du cas où k_m est une modalité associée à la variable qualitative définie comme suit. Soit $\mathcal{K} = \prod_{h=1}^H \mathcal{K}_h$, où chaque $\mathcal{K}_h = \{y_{h1}, \dots, y_{hM(h)}\}$ correspond à un caractère qualitatif à M_h modalités, et $\kappa : \Omega \mapsto \mathcal{K}$ est une va multivariée (ie H-variée) donnée. L'application :

$$(1) \quad \omega \in \Omega \mapsto \kappa(\omega) \in \mathcal{K}$$

définie par κ est supposée être une **application surjective** (si Ω est fini, $\text{Card } \mathcal{K} = \prod_{h=1}^H M_h \leq \text{Card } \Omega$), ie il existe au moins un individu ayant une caractéristique donnée $k \in \mathcal{K}$.

(v) Dans le cas d'un caractère qualitatif κ , il est possible d'associer à chaque élément (individu, objet, etc) ω une **variable qualitative** privilégiée $d_m(\omega)$ tq :

$$(2) \quad d_m(\omega) = \begin{cases} 0 & \text{si } \kappa(\omega) \neq k_m, \\ 1 & \text{si } \kappa(\omega) = k_m, \end{cases}$$

ie la **variable indicatrice** de l'**événement** $[\kappa = k_m]$. Ceci revient à définir un **codage** particulier sur $\Omega \times \mathcal{K}$.

(vi) En **analyse des données**, un **codage (numérique)** de $\mathcal{K}$ par un ensemble numérique $\mathcal{X}$ est une **application bimesurable** $c : \mathcal{K} \mapsto \mathcal{X}$ tq il existe une **va** surjective ξ vérifiant : $\xi = c \circ \kappa$.

(vii) L'« opposition » souvent faite entre qualitatif et quantitatif n'est pas un obstacle au champ d'application de la Statistique. Ainsi, si $\mathcal{K}$ représente l'étendue (supposée continue) du spectre lumineux et Ω l'ensemble des fleurs d'une serre, κ peut associer à chaque fleur, au lieu d'une fréquence (ou d'une longueur d'onde) déterminée $x_m \in \mathcal{X}$ (ensemble numérique) un nom de couleur $k_m \in \mathcal{K}$, défini en partitionnant $\mathcal{X}$ de façon suffisamment fine. On peut suivre la démarche inverse si l'on dispose, au départ, d'une application $\kappa : \Omega \mapsto \mathcal{K}$.

(viii) Du point de vue terminologique, il semble préférable de se référer à la distinction générale entre variable quantitative (ou numérique), variable qualitative ou encore variable morphologique.

Lorsqu'il s'agit d'une **variable qualitative**, un caractère est aussi appelé **attribut**.

Cependant, la terminologie n'est pas toujours précise. Les termes « caractère » ou « attribut », qui possèdent une connotation plutôt qualitative, ne seront pas associés, dans ce dictionnaire, à des variables numériques mais, le cas échéant, à des variables qualitatives.