CLASSIFICATION DES VARIABLES (C1)
(13 / 10 / 2019, © Monfort, Dicostat2005, 2005-2019)

Pour préciser le mode d'**observation** d'un **phénomène** donné, chacun des **descripteurs** utilisés représente un « objet » ou « concept » général de la **science** : ce concept sert à « décrire », de façon interprétable, chaque élément d'un **ensemble** (eg une **unité statistique**) ou une **situation statistique**. Le concept statistique correspondant à ce descripteur est celui de **variable** ou celui de **variable aléatoire**. : en effet, la valeur observée (ou « prise ») par cette variable varie généralement d'une unité à une autre, l'aléas considéré pouvant être :

(a) « intrinsèque » : **variabilité** propre à la **population** étudiée ;

(b) ou « extrinsèque » : mode d'observation faisant usage de procédés probabilistes (cf **schéma probabiliste**).

(i) Trois notions de variables peuvent être distinguées :

(a) variable **mathématique** ;

(c) **variable aléatoire** (**va**), ie variable faisant l'objet du **calcul des probabilités** ;

(c) **variable statistique**, ie variable faisant l'objet de la **Statistique descriptive**. Il s'agit d'un **caractère** observé (ie décrit ou mesuré), ou observable, sur une unité statistique ;

On ne distingue pas toujours entre variable aléatoire et variable statistique : une variable statistique ne fait pas nécessairement référence à un **espace probabilisé** (ceci est le cas en Statistique descriptive). Inversement, une variable aléatoire constitue une façon commode de traiter statistiquement un « caractère » statistique jouant le rôle de descripteur.

(ii) Une variable aléatoire (ou une **statistique**) peut ainsi être :

(a) une **variable qualitative**, parfois appelée **caractère** ou **attribut** ;

(b) une **variable quantitative**, ie une grandeur numérique : une telle variable est à valeurs dans un ensemble numérique usuel : ensemble « discret » ($\mathbb{N}_n$, $\mathbb{Z}_{mn}$, $\mathbf{N}$, $\mathbf{Z}$, $\mathbf{D}$, $\mathbf{Q}$) ou « continu » ($\mathbf{R}$, $\mathbf{C}$), ou dans leurs puissances cartésiennes.

(c) une variable **variable morphologique** : une telle variable peut être distinguée des précédentes, car elle possède une double nature. D'un côté, elle peut se représenter numériquement (ie à l'aide d'équations ou inéquations) : eg **formes** diverses (figures géométriques de la **géométrie stochastique**), ou formes relevant de la **Statistique géométrique**. De l'autre, elle peut correspondre à une variable qualitative, dans la mesure où les formes de première nature peuvent faire l'objet de **classification**.

(iii) La **théorie du codage** permet de définir des variables quantitatives à partir de variables qualitatives. On peut en faire de même lorsque les variables qualitatives

consistent en des descriptions basées sur des variables numériques : ainsi, la variable « couleur des yeux » à partir des **fréquences** du spectre lumineux ; la variable « morphologie externe du corps humain », ou « surface du corps humain » à partir d'une classification des équations définissant cette morphologie ou cette surface (cf **exemple de l'introduction**) ; etc.

(iv) Inversement, on peut définir des variables qualitatives à partir de variables quantitatives. Ceci peut résulter (a) d'une **discrétisation** de lois, (b) d'un partitionnement des espaces numériques considérés en « classes » disjointes (cf **partition**), ou encore d'une **classification automatique** appliquée à des données, numériques ou non.

(v) Dans un **modèle statistique** (en particulier, un **modèle de régression**) ou encore en **analyse multidimensionnelle** (notamment, l'**analyse des données**), on est souvent conduit à « combiner » les variables utilisées (qualitatives ou numériques) :

(a) selon leurs propriétés (variables endogènes et variables exogènes). Ainsi, on parle de **modèle quantitatif** (ou **modèle numérique**), de **modèle qualitatif**, de **modèle mixte** (ie comportant des **va** quantitatives et des va qualitatives) ;

(b) selon leur nombre : « liste » de variables décrivant le phénomène, notamment en relation avec le nombre d'observations (cf **degré de liberté**). Ainsi, on définit des variables simples (ou univariées, ou unidimensionnelles) et des variables multiples (ou multivariées, ou multidimensionnelles, ie à plusieurs dimensions) : les premières sont à valeurs dans un ensemble donné (en général, « simple » ou de nature « scalaire »), les secondes dans un produit cartésien d'ensembles du type précédent (eg un **espace vectoriel** dont la dimension est égale au nombre d'espaces facteurs).

(vi) Enfin, une variable numérique peut être une **variable discrète** ou une **variable continue**, voire une **variable mixte**.

Une variable discrète (eg une variable qualitative) est parfois elle-même une variable codée à partir d'une (ou plusieurs) variables qualitative(s) (cf **codage d'un modèle statistique**, **code**).