DONNÉE (O)
(21 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

(i) Le mot **donnée** est un terme général, plus ou moins vague, pouvant recevoir une signification dépendant du contexte :

(a) **ensemble des « données » relatives à un phénomène** étudié : état d'avancement d'un corps de connaissance, théorie du phénomène, données au sens plus restreint suivant ;

(b) **ensemble des observations portant sur des variables statistiques** considérées comme pertinentes vis à vis du problème posé, ou simplement portant sur les variables disponibles. Ce type de données se présente généralement soit sous forme de données détaillées (ou données individuelles), soit sous forme de données agrégées (cf **agrégation**, **tableau statistique**), établies selon des classifications croisées, ou selon des classifications emboîtées, équilibrées ou non, avec ou sans **lacunes** (« trous » ou observations manquantes). Si les données résultent de comptages, on les appelle **données de dénombrement** ;

(c) **ensemble des observations faisant l'objet d'une analyse des données**.

(ii) Les données statistiques sont souvent qualifiées comme les variables dont elles représentent les valeurs (observations ou mesures répétées des variables sur des unités) :

(a) données quantitatives (ou numériques) ou qualitatives (ou d'attributs) (cf **variable qualitative**, **variable quantitative**) ;

(b) données continues ou discrètes (cf **variable continue**, **variable discrète**), etc.

(iv) Une donnée, considérée comme observation d'une variable, est à distinguer d'une « **inconnue** » qui est, le plus souvent :

(a) soit une variable inobservable : **perturbation aléatoire** d'un **modèle de régression** ou d'un **modèle d'interdépendance**, **facteur** d'une **analyse en facteurs communs et spécifiques**, classe latente, etc) (cf **structure latente**) ;

(b) soit un **paramètre d'intérêt** pour le statisticien (cf **paramètre**).

Cette disctinction diffère de celle, mathématique, entre donnée et inconnue (eg dans un problème, dans la résolution d'un système d'équations, etc), qui intervient aussi en **Statistique**.

(v) Une distinction utile se rapporte au éléments suivants, souvent produits par un **système statistique** (cf **production statistique**) :

(a) **donnée individuelle** : c'est l'observation ξ (a) d'une **variable** quelconque ξ, appelée **attribut** (**variable numérique**, **variable qualitative** ou **variable**

morphologique), relative à une **unité statistique** a $\in$ A, **échantillon** extrait d'un **ensemble** donné Ω (**population**) ;

(b) **donnée temporelle** : une telle donnée peut être individuelle x_{at} (eg « trajectoire biologique ou physiologique », ou encore « trajectoire sociale » d'une personne physique a au cours du temps) (cf **monographie**), mais elle peut aussi résulter d'une **agrégation** $x_A(t) = \Sigma_{a \in A} x_{at}$ de données individuelles (au sens précédent) (eg prévisions météorologiques, évolutions démographiques, séries d'emploi ou d'indices de prix, etc). Ici, t $\in$ T désigne un instant élémentaire de la période de **temps** T considérée ;

(c) **donnée spatiale** : comme les deux précédents types de données, une donnée de cette nature peut être individuelle x_{as} (eg IMC de personnes physiques identifiables tirées au hasard dans diverses communes) ou « groupée » selon $x_A(s) = \Sigma_{a \in A} x_{as}$ (eg répartition d'une population ou du PIB par région ou par département, etc). Ici, s $\in$ S désigne une « zone » élémentaire de l'**espace** S considéré (eg S = $\mathbf{R}^3$).

Ainsi, la plupart des données, individuelles ou agrégées, peuvent être (au moins mentalement) indexées par le triplet (a, t, s) $\in$ A x T x S (cf aussi **niveau, répartition, évolution**). Dans certains cas, A = Ω (recensement).

De nombreuses autres données dérivent des précédentes : taux d'évolution, **proportions** ou **rapports** divers, etc.