POPULATION (B1, C4, M)
(02 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

Le mot **population**, d'origine démographique, est un terme plus « technique », très général, et quasiment « générique » : il est utilisé en **Statistique** avec la même signification que le mot « **ensemble** » en mathématique.

Cependant, le **contexte statistique** est généralement probabiliste : aussi, une population doit faire l'objet d'une structuration afin de permettre l'analyse statistique (**plan d'expérience**, **plan de sondage**, **modélisation**).

(i) On appelle **population** un ensemble dont chaque élément est appelé **unité statistique**, ou « **individu** ». Un tel élément peut être :

(a) physique : un objet (particule, rayonnement) ;

(b) biologie : un microbe (virus, bactérie), une cellule vivante plus complexe ;

(c) écologie : un membre d'une espèce animale ou végétale ;

(d) psychologie : une personne physique (sujet) ;

(e) sociologie : une personne physique, une personne morale (entreprise, groupe de sociétés, administration, nation constituée ou non, etc).

Une population peut cependant concerner des objets mathématiques tq : nombre, fonction (eg **application continue**), surface ou variété, etc (cf eg **forme**, **géométrie stochastique**, **probabilité géométrique**).

(ii) Selon le contexte, une population est aussi appelée de façon différente :

(a) **calcul des probabilités** : **ensemble des épreuves (élémentaires)**, ou **ensemble des évènements (élémentaires)**. Tout élément $\omega \in \Omega$ est alors appelé **épreuve** ou **événement** (élémentaire) ;

(b) **théorie de la mesure**, **Statistique** : **ensemble fondamental**. Tout élément $\omega \in \Omega$ est alors appelé **unité statistique** ou **unité élémentaire** ;

(c) **théorie des sondages** : **univers**. Tout élément $\omega \in \Omega$ est alors appelé **unité de sondage**, ou parfois **« astre »**, etc ;

(d) **théorie des plans d'expérience** : **dispositif expérimental**. Tout élément $\omega \in \Omega$ est alors appelé **unité expérimentale**.

(iii) On note ainsi souvent une population à l'aide du symbole Ω :

(a) si Card Ω = M est fini, on parle de **population finie**, cas généralement analysé en **théorie des sondages**. L'entier $M \geq 1$ est aussi appelé **effectif**, ou **taille**, de la population ;

(b) si Card $\Omega = \aleph_0$ (**puissance « aleph zéro »** du dénombrable), on parle de **population dénombrable** (en bijection avec **N** ou l'une de ses parties) ;

(c) par extension, on est souvent amené à considérer des populations dont la puissance est supérieure à celle du dénombrable : eg l'ensemble des **trajectoires** d'un **processus** peut posséder la puissance du continu (en bijection avec **R** ou l'une de ses parties).

(iv) Si $\mathcal{T}$ est une **tribu de parties** de Ω et si P est une **mesure de probabilité** définie sur $\mathcal{T}$, le triplet $(\Omega, \mathcal{T}, P)$ est appelé **espace probabilisé**.

$\mathcal{T}$ est souvent appelé **ensemble des épreuves (complexes)**, ou **ensemble des évènements (complexes)** (**calcul des probabilités**). Chaque **partie mesurable** $A \in \mathcal{T}$ est appelée **épreuve** ou **événement** (complexe).

On dit souvent, par extension, que la **probabilité** P elle-même représente la **population** étudiée.

(v) Si l'on « observe » (mesure, description), sur chaque individu $\omega \in \Omega$, une **grandeur** notée $y = \eta(\omega)$ appartenant à un ensemble d'**observations** $\mathcal{Y}$, on peut définir l'**espace probabilisé image** de l'espace précédent par l'application $\eta : \Omega \mapsto \mathcal{Y}$. Cette application est, de façon générale, une **variable** souvent appelée **attribut**, **caractère**, **descripteur** ou encore **facteur**. Chaque « valeur » $y = \eta(\omega)$ est aussi appelée **donnée** ou **observation** (s'il y a **observabilité**), ou **mesure** effectuée sur l'unité ω.

On dit aussi que $\mathcal{Y}$ est la **population** étudiée (ensemble des valeurs) ; de même, on dit aussi que la **mesure image** P^η de P par η est une « **population** » (ensemble des masses de probabilité affectées aux valeurs précédentes).

Ces notions procèdent ainsi d'une **distinction importante** entre :

(a) la population Ω constituée d'unités statistiques ;

(b) la « population » $\mathcal{Y}$ constituée d'**observations** (mesures, descriptions) effectuées sur ces unités et relatives à diverses **variables** qui les décrivent et qui peuvent être observées (cf **observabilité**).

(vi) L'étude des **caractéristiques d'une population** Ω (resp $\mathcal{Y}$) se fait généralement à l'aide d'un sous-ensemble $A \subset \Omega$, appelé **sous-population** de Ω (souvent, $A \in \mathcal{T}$, tribu de parties de Ω).

En **théorie des sondages**, on définit plutôt une **partie** $A \in \Omega^N$, ie $A \subset \mathcal{P}(\Omega^N)$. Autrement dit, $A = \{a_1, ..., a_N\}$ est appelée **« échantillon » extrait** de la population Ω, et on la note aussi $A = (a_1, ..., a_N)$.

Cette extraction peut s'effectuer de deux façons principales :

(a) si aucun **schéma probabiliste** (ou mécanisme probabiliste) (connu) ne gouverne cette extraction, on parle de **sondage non aléatoire** (a priori). On peut cependant parfois considérer un tel sondage comme un sondage aléatoire a posteriori ;

(b) si les unités a_n de A sont tirées dans Ω selon une **loi de probabilité** Π donnée, définie sur $\mathscr{P}(\Omega^N)$, on parle de **sondage aléatoire**, ou de **sondage probabiliste** (cf **probabilité d'inclusion** in **estimateur de HORWITZ-THOMPSON**).

(vii) Il importe de distinguer entre deux « distributions » :

(a) la « **description** » de la population Ω à travers une **variable observable** η (numérique ou non) : cette description implique une répartition des valeurs de η entre les éléments ω de Ω, dont se déduit une « distribution de fréquences » de ces valeurs. Cette distribution est alors interprétée comme « loi de probabilité théorique ». Elle est notée P^η ou $\mathscr{L}(\eta)$;

(b) celle définissant le choix des unités finalement observées, ie le « **tirage** » de ces unités : la **lp** Π est la distribution qui détermine ce choix.

On distingue ainsi entre :

(a) l'**espace probabilisé de base** (Ω, $\mathscr{T}$, P), ou **espace sous-jacent** ;

(b) l'**espace probabilisé image** (a priori observable) ($\mathscr{Y}$, $\mathscr{C}$, P^η) (**espace d'observation**) ;

(c) l'**espace de sondage** (Ω^N, $\mathscr{P}(\Omega^N)$, Π) (cf **modèle de sondage**).

C'est ce dernier espace qui génère le mécanisme aléatoire selon lequel s'effectue le tirage des unités, donc la **va** « **échantillon** » $A = (a_1, ..., a_N)$ et, par suite, la va « observée » $y = (y_1, ..., y_N)$, avec $y_n = \eta(a_n)$, $\forall\ n \in N_N^*$ (cf **espace d'échantillonnage**).

(viii) Un tirage non aléatoire peut être considéré comme un cas particulier de tirage aléatoire dans lequel $\Pi = \delta(A_0)$ (**loi de DIRAC** en un « point » A_0), le N-uple A_0 étant donné a priori (d'où une optique bayésienne possible), ou encore comme un tirage dans une **superpopulation** (autre optique bayésienne possible) (cf **principe bayésien**).

(ix) La théorie des sondages, ou la théorie de l'**échantillonnage**, ont pour objet l'étude des relations entre :

(a) ce que l'on observe au niveau de l'échantillon ;

(b) et ce que l'on n'observe pas au niveau de la population (**inférence statistique**).

Le moyen utilisé est l' « intermédiaire » que constitue le phénomène aléatoire (**schéma probabiliste**) décrit par la probabilité Π.