PROCESSUS GÉNÉRATEUR DE DONNÉES (E3, N2)
(05 / 08 / 2020, © Monfort, Dicostat2005, 2005-2020)

(i) De façon générale, on appelle **processus générateur de données**, ou **processus de génération des données**, (pgd) (en anglais « *data generating process* » (dgp)), un **modèle statistique** supposé avoir engendré les **données** relatives à un **phénomène** considéré.

Il existe, en effet, trois modes de connaissance d'une **information**, ou « **donnée** », associée à l'**observation** d'un phénomène (cf **production statistique**) :

(a) un **mode non statistique**, qui consiste à utiliser les informations disponibles, quelle que soit la façon dont elles ont été « construites ». Cette façon n'est pas toujours connue du **statisticien** ;

(b) deux **modes statistiques** :

$(b)_1$ par **expérimentation** (cf **plan d'expérience**) ;

$(b)_2$ par **sondage** (cf **plan de sondage**).

(ii) Le mode non statistique concerne généralement des informations à caractère « historique » : eg données climatiques, écologiques, sociologiques (démographie).

Ces informations sont souvent produites par un organisme (administration publique ou privée). Ainsi :

(a) diverses chroniques anciennes (eg du Moyen Age) relatent les lieux et dates concernant des particularités climatiques (sécheresse exceptionnelle, hivers particulièrement rigoureux, pluviométrie rare, etc), des séismes, des épidémies, etc ;

(b) de même, des recensements du royaume de France ont été effectués (eg Saugrain l'Ainé en 1709 ou 1729, etc). Ceux-ci dénombrent, par généralité, baillage, sénéchaussée et paroisse, le nombre de « feux », ou « foyers ». Cette notion est à rapprocher des actuels « foyers fiscaux » ou « ménages » (comptabilité nationale). Ces recensements s'associent aux pouvoirs régaliens anciens : levage de troupes, prélèvement d'impôts.

Les chroniques ou autres investigations, de nature non statistique, peuvent ainsi permettre de constituer des bases pour compléter, rétrospectivement, les informations disponibles actuellement de façon systématique, mais sur des périodes relativement courtes (un siècle ou deux) : climatologie, physique du globe, santé publique, démographie. Ainsi, par **prévision** « arrière » (rétropolation), le Bureau de référence sur la population (PRB, www.prb.org) a pu estimer, au niveau mondial, la population des êtres humains ayant existé sur Terre (environ 108 milliards en 2011, soit environ 15 fois la population d'alors).

Les informations antérieures sont souvent hétérogènes pr à celles actuellement élaborées. Il est parfois possible, eg par **modélisation**, de donner un contenu cohérent à l'ensemble des données : rétrospectives anciennes ou plus récentes.

(iii) Les méthodes de **génération de nombres au hasard** font aussi partie des procédés par lesquels on construit des données (cf **échantillon artificiel**).

(iv) Enfin, les capacités de calcul et de stockage informatiques, ainsi que l'existence de **réseaux** d'information de grande taille, permettent la constitution de **grandes bases de données** : ces « gisements » d'information peuvent alors être analysés statistiquement afin d'apprécier certains phénomènes ou de vérifier certaines idées relatives à ces derniers.