MOYENNE (C5, F3, H3)
(31 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

Une **moyenne** est un **moment** particulier, donc une **caractéristique légale** particulière très utilisée.

(i) Généralement, une **moyenne** constitue un « indicateur » de **centralité** : cette valeur est considérée comme « typique », ou « représentative » :

(a) soit d'un **ensemble** donné (**population** ou **échantillon**) constitué de diverses **unités statistiques** généralement comparables ;

(b) soit de la **variable aléatoire** ou de la **loi de probabilité** décrivant cet ensemble (population, échantillon). Cette variable est essentiellement une **variable numérique**.

Autrement dit, l'ensemble des **unités statistiques** (population ou échantillon d'individus, **matériel expérimental**) est « observé » à travers une **variable** donnée, qui s'interprète comme une **variable aléatoire** : **attribut**, **caractère**, etc.

La moyenne est alors considérée comme une **« valeur caractéristique »**, ou un **« résumé »**, de l'ensemble des valeurs prises par cette **va**. Cette caractéristique joue donc un rôle de **valeur centrale** : cf **centrage**, **centralité**, **loi non centrale**, **moment algébrique**, **partie centrale**, **valeur typique de FRÉCHET**.

Une moyenne est de même nature que la variable observée sur l'ensemble considéré : en effet, les deux notions sont de même dimension pr à l'**unité de mesure** (cf **échelle de mesure**) et l'une peut se différencier ou se rapporter à l'autre (cf **écart**, **rapport**).

(ii) Par définition, la **moyenne théorique** est l'**espérance mathématique** de la **variable aléatoire** ou de la **loi de probabilité** décrivant la population.

Souvent, un **modèle statistique** admet une moyenne pour **paramètre**. Ainsi en est-il du **modèle gaussien**, dont la famille des lois $(P_\theta^\xi)_{\theta \in \Theta}$ est de la forme $(\mathcal{N}(\mu, \Sigma))_{(\mu, \Sigma)}$, avec $\theta = (\mu, \Sigma)$ et $\theta \in \mathbf{R}^K \times M_K^{++}(\mathbf{R})$, où $M_K^{++}(\mathbf{R})$ désigne le cône des **matrices définies positives**.

(iii) A la notion de moyenne théorique précédente correspond celle de **moyenne empirique**. Cette dernière est calculée à partir d'un **échantillon**, en remplaçant dans les calculs la loi de probabilité « théorique » par la **loi empirique** (cf **statistique naturelle**). Cet échantillon n'est pas nécessairement un **échantillon indépendant** ou un **échantillon équidistribué**.

En particulier, un outils souvent utilisé dans certains calculs algébriques (eg **modèle linéaire**, **modèle d'interdépendance linéaire**) est la **matrice de centrage par rapport à la moyenne** (empirique).

(iv) Comme toute **statistique**, une moyenne est calculable :

(a) à l'aide de tout, ou partie, des valeurs observables (cf **observabilité**) ;

(b) sans altérations ou avec altérations.

On peut alors distinguer plusieurs types de moyenne, selon que l'on prend en compte :

(a) le **mode de calcul** : **moyenne arithmétique**, **moyenne géométrique**, **moyenne harmonique**, **moyenne exponentielle**, **moyenne potentielle** ;

(b) l'**objet du calcul** (ou calculs spécifiques) : **moyenne mobile** (séries temporelles), **moyenne dans L^p**, **moyenne de CESARO** ;

(c) une **« valeur de référence »**, ou **« valeur de centrage »** : on distingue alors entre **moyenne non centrée** (avec valeur de référence « nulle ») et **moyenne centrée** ;

(d) la **multivariabilité** ou la **dimensionnalité** de la **loi de probabilité** de la variable considérée : loi monovariée ou **loi multivariée**, loi monodimensionnelle ou **loi multidimensionnelle**. Dans le cas multidimentionnel, on distingue les notions de **moyenne marginale** et de **moyenne conditionnelle** (cf **caractéristique conditionnelle**, **espérance conditionnelle**) ;

(e) d'éventuelles **modifications** apportées au mode de calcul : **moyenne équilibrée**, **moyenne d'ensemble** ou moyenne partielle (calcul sur « champ » restreint : sous-population, strate, etc) ;

(f) **divers « accidents »** pouvant affecter le calcul : **moyenne censurée**, moyenne calculée en présence de **lacunes** (cf aussi **observation manquante**) ;

(g) une **éventuelle « altération »** de la loi de probabilité (cf **troncature**, **mélange de lois**, **loi multimodale**).

(v) De façon générale, et sans autre précision, le terme de « moyenne » fait (par défaut) référence à la **moyenne arithmétique** (simple ou pondérée). Dépendant linéairement des observations, ce type de moyenne est un indicateur de centralité facile à calculer, ce qui explique sa fréquente utilisation, compte tenu de la simplicité de son interprétation.

(vi) Cependant, la moyenne n'est pas toujours un concept pertinent. C'est pourquoi il existe des **concepts alternatifs** à la moyenne : **médiane**, **mode**, **quantiles**, etc. Cette absence de pertinence peut tenir à trois causes principales :

(a) **inexistence** : le moment (théorique) qu'est l'**espérance** (moyenne théorique) n'existe pas (cas de la **loi de CAUCHY**). On ne peut donc pas la calculer, alors même que la moyenne empirique correspondante peut toujours être « formellement » calculée ;

(b) **forme de la loi** (cf **coefficient de forme**, **forme d'une loi**). Une distribution asymétrique (unimodale) (cf **asymétrie**, **loi symétrique**, **loi unimodale**) correspond souvent à une **variable positive** (masse d'un élément, nombre d'individus, taille ou poids d'un individu, revenu d'activité économique, etc). Calculer la moyenne d'une telle distribution est sans doute moins pertinent qu'en calculer eg le **mode** ou la **médiane**

(selon le contexte), notamment si cette loi est dissymétrique (cf **loi symétrique**), et notamment une **loi oblique** (cf **oblicité**) ;

(c) **altération de la loi**. Ainsi :

$(c)_1$ en présence d'une **loi multimodale**, le calcul d'une moyenne arithmétique est dénué de sens : il vaut mieux procéder à la **dissection du mélange de lois** sous-jacent et calculer les moyennes des sous-populations correspondantes (sous-moyennes), même si la moyenne d'ensemble pondérée de ces sous-moyennes est (par définition) égale à la moyenne initiale ;

$(c)_2$ de même, si la loi considérée est tronquée (cf **troncature**), la moyenne (empirique) « formellement » calculée à partir d'un échantillon n'a pas nécessairement un sens clair, ou concret. L'échantillon issu d'une loi tronquée peut parfois être considéré comme un échantillon censuré (cf **censure**).

(d) **loi qualitative**, ie loi d'une **variable qualitative** (**attribut**) : si cette variable est « valuée », sa moyenne se définit comme pour un moment algébrique quelconque. Dans les autres cas, deux approches sont concevables :

$(d)_1$ codage numérique : à chaque **modalité** de la variable est associée une valeur numérique : soit par **codage** ;

(d)2 calcul d'un **score statistique**.

La moyenne est alors calculée à l'aide des valeurs obtenues. Cette approche peut se justifier, par « remontée », pour une variable qualitative déduite d'une variable numérique (eg couleurs des yeux d'individus et fréquences du spectre lumineux)