PROPORTION (C5, F3, K1)
(21 / 11 / 2020, © Monfort, Dicostat2005, 2005-2020)

Le **statisticien** utilise souvent les notions de **proportion** ou de **rapport** : **indices** temporels ou spatiaux, **coefficients** divers, règles de trois, pourcentages, etc (cf **quotient**, **estimateur par règle de trois**).

Ces outils numériques sont ainsi liés aux notions de **structure** ou de **relativité**, ou encore de **relativisme**.

En particulier, en **théorie des sondages**, il est souvent utile d'estimer le nombre d'**unités statistiques** (individus) possédant un descripteur (**attribut** ou **caractère**) donné, ou la proportion de celles-ci relativement à un **ensemble** de référence Ω.

(i) Soit $\Omega = \{\omega_1 ,..., \omega_M\}$ un **ensemble** fini constitué d'**unités statistiques** ω_m (eg une **population** d'individus), $(\mathcal{Y}, \mathcal{G})$ un **espace d'observation** quelconque (ie numérique ou non) et $\eta : \Omega \mapsto \mathcal{Y}$ un **caractère statistique** (quantitatif ou qualitatif) défini sur les unités de Ω (cf **variable qualitative**, **variable quantitative**). On note P_M (ou P^η) la « distribution » de η dans Ω :

$$(1) \qquad P_M = M^{-1} . \Sigma_{m=1}^{M} \delta(Y_m),$$

où $Y_m = \eta (\omega_m)$ est, $\forall$ m $\in N_M^*$, la valeur du caractère η observée sur ω_m , et $\delta(Y_m) = \delta(\eta (\omega_m))$ désigne la **masse de DIRAC** placée en ω_m .

Soit C $\in \mathcal{G}$ une **partie** donnée de Ω. La « **probabilité** » :

$$(2) \qquad p = P_M ([\eta \in C])$$

définit la **proportion** des unités $\omega_m \in \Omega$ possédant le caractère C défini par deux modalités $\{c_1 , c_2\}$ selon :

$$(3) \qquad \begin{array}{l} \eta (\omega_m) \in C \Rightarrow \omega_m \text{ possède la modalité } c_1 , \\ \eta (\omega_m) \notin C \Rightarrow \omega_m \text{ possède la modalité } c_2 . \end{array}$$

(ii) Si l'on « code » numériquement la variable qualitative η à l'aide de la **variable indicatrice** $\mathbf{1}_C : \Omega \mapsto N_1$ (cf **codage**, **variable de CORNFIELD**), alors :

$$(4) \qquad p = M^{-1} . \Sigma_{m=1}^{M} \mathbf{1}_C (\omega_m)$$ (**moyenne** des indicatrices dans la population).

(iii) Enfin, si A est un N-**échantillon aléatoire** extrait de Ω selon un **plan de sondage** Π, on note $A = \{a_1 ,..., a_N\}$ et $y_n = \eta (a_n)$, $\forall$ n $\in N_N^*$. On utilise alors souvent l'**estimateur** suivant du **paramètre d'intérêt** p :

(5) $T_N''' = N^{-1} \Sigma_{n=1}^{N} \mathbf{1}_C (a_n)$ (moyenne des indicatrices dans l'échantillon).

Cet estimateur, aussi noté f_N , est appelé **fréquence empirique**. En pratique, il se calcule directement à partir d'observations y_n tq $y_n = \mathbf{1}_C (a_n)$. Par suite :

(6) $T_N''' = \bar{y}_N = N^{-1} . \Sigma_{n=1}^{N} y_n$.

Une proportion n'est ainsi autre chose qu'une moyenne (arithmétique) particulière.