TABLEAU STATISTIQUE (C1, F, G, J, K1)
(23 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

Un **tableau statistique** constitue le « matériau » de base de nombreux travaux statistiques : son contenu et l'interprétation concrète qu'on peut lui donner dépendent de la nature des **observations** collectées et sont très variés.

De façon générale, un tableau statistique est un couple formé :

(a) d'une **matrice**, éventuellement multidimensionnelle (cf tenseur). Cette matrice comporte divers « descripteurs », qui peuvent être des symboles divers (**variables qualitatives**) ou des nombres (**variables quantitatives**) ;

(b) de l'ensemble des **indices** (ou multi-indice) repérant les éléments de cette matrice (ou « cases » du tableau). A ces indices peuvent être associées les modalités de divers **codes** (ie variables qualitatives).

Ainsi, un tableau à deux dimensions peut décrire une variable répartie selon divers « croisements », eg (sociologie) :

(a) (économie) revenu disponible brut (nombre décimal) distribué selon le croisement entre I catégories sociales (modalités A, B, ..., Z), numérotées selon les indices i = 1 ,..., I, et J zones géographiques (modalités 01, 02, ..., 99), numérotées selon les indices j = 1 ,..., J ;

(b) (démographie) nombre moyen d'enfants par type de ménage (nombre entier) ;

(c) (politologie) la tendance politique dominante (eg code {bleu, vert, rouge}).

Un tableau statistique peut donc contenir :

(a) des nombres entiers (cf **tableau de contingence**) ;

(b) des nombres quelconques : nombres réels (éléments de **R**), nombres rationnels (éléments de **Q**) ou, le plus souvent, nombres décimaux (éléments de **D** : ensemble des nombres rationnels comportant un nombre fini de chiffres après la virgule) ;

(c) des signes variés (tels que + , = , - , . , etc) qui peuvent souvent s'associer à des **caractères statistiques** qualitatifs ou à des **variables qualitatives** (cf aussi **codage**).

(i) Soit $\{\Omega, \mathcal{T}, P)$ un espace probabilisé, $((\mathcal{X}_\alpha , \mathcal{B}_\alpha))_{\alpha=1,...,K}$. une suite finie d'**espaces mesurables** (**espaces d'observation**, le plus souvent) et $\xi = (\xi_1 ,..., \xi_k)$, une suite finie de **va** quelconques $\xi_\alpha : \Omega \mapsto \mathcal{X}_\alpha$, qui s'interprètent comme décrivant les éléments ω de l'ensemble fondamental Ω (individus d'une **population**). Par suite, le

k-uple aléatoire $\xi = (\xi_1 ,..., \xi_k) : \Omega \mapsto \mathcal{X}$, où $\mathcal{X} = \prod_{\alpha=1}^{k} \mathcal{X}_\alpha$, ainsi défini correspond à k **caractères** pouvant être observés sur les éléments $\omega \in \Omega$.

A chaque espace $(\mathcal{X}_\alpha , \mathcal{B}_\alpha)$ on associe une **partition** $\Pi_\alpha = (B_{\alpha i(\alpha)})_{i(\alpha)}$ (où $i \in \{1 ,... m_\alpha\}$) (ou simplement une **suite** finie de **parties** disjointes $B_{\alpha i(\alpha)}$ de $\mathcal{X}_\alpha$) et l'on note $i_\alpha \in \mathcal{I}_\alpha = \{1 ,..., m_\alpha\}$, $I = (i_1 ,..., i_k) \in \mathcal{I} = \prod_{\alpha=1}^{k} \mathcal{I}_\alpha$ (multi-**indice** ou k-indice).

On observe alors N éléments $a_1 ,..., a_N$ de Ω, constituant un **échantillon** $A \subset \Omega^N$. On peut définir le nombre d'éléments a_n qui appartiennent à tous les $B_{\alpha i(\alpha)}$ ($\alpha = 1 ,..., k$) (cf **tableau de contingence**).

Par ailleurs, on définit, pour tout k-uple de parties $(B_{1i(1)} ,..., B_{ki(k)}) \in \prod_{\alpha=1}^{k} \Pi_\alpha$, une **application** $\zeta_I : \prod_{\alpha=1}^{k} \Pi_\alpha \mapsto \mathcal{Z}_I$ où $I = (i_1 ,..., i_k)$ et où, pour tout $I \in \mathcal{I}$, $\mathcal{Z}_I$ est un ensemble définissant la « représentation » des va dans le tableau :

(a) si les variables sont numériques (**variables quantitatives**), il peut s'agir eg de totaux ou d'agrégats (cf **agrégation**) ;

(b) si les variables sont non numériques (**variables qualitatives**), il peut s'agir de symboles divers.

En général, card $\mathcal{Z}_I = 1$ et l'on note simplement z_I au lieu de $\mathcal{Z}_I$.

On appelle alors **tableau statistique** la donnée de la suite $T = (z_I)_{I \in \mathcal{I}}$ ainsi définie et des modalités des variables de croisement. Ce tableau est dit **multivarié (ou k-varié)**, ou encore **multidimensionnel (ou k-dimensionnel)** car chaque élément $\omega \in \Omega$ est « repéré » à partir de k variables ξ_α et que le k-uple I indicie les z_I .

(ii) Lorsque $T = (z_I)_{I \in \mathcal{I}}$ est constitué de nombres appartenant à un corps **K** (eg **K** = **R** ou **K** = **C**), on peut associer T à une forme k-linéaire (cf **forme multilinéaire**).

(iii) Un cas particulier important de tableau statistique est celui du **tableau à deux dimensions** (k = 2). On note alors souvent z_{ij} (ou z_{rc}) au lieu de $z_{i(1)i(2)}$, où $i \in \{1 ,..., I\}$ et $j \in \{1,..., J\}$ (ou bien $r \in \{1 ,..., R\}$ et $c \in \{1 ,..., C\}$). Par suite, $T = (z_{ij})_{(i,j)}$ est un tableau statistique qui s'associe naturellement à une **matrice** lorsque les z_{ij} prennent leurs valeurs dans un corps **K** (souvent, **K** = **R** ou **C**).

Tableau statistique numérique à deux dimensions
(les critères de croisement ne sont pas représentés)

z_{11}	…	z_{1j}	…	z_{1J}	$z_{1.}$
…	…		…	…	…
z_{i1}	…	z_{ij}	…	z_{iJ}	$z_{i.}$
…	…	…	…	…	…
z_{I1}	…	z_{Ij}	…	z_{IJ}	$z_{I.}$
$z_{.1}$	…	$z_{.j}$	…	$z_{.J}$	$z_{..}$

On appelle alors :

(a) **première marge**, ou **marge de droite**, la colonne T e_J formée des I éléments $z_{i.} = \sum_{j=1}^{J} z_{ij}$;

(b) **deuxième marge**, ou **marge du bas**, la ligne e_I' T formée des J éléments $z_{.j} = \sum_{i=1}^{I} z_{ij}$;

(c) **total général** le scalaire $z.. = e_I'$ T $e_J = \sum_{i=1}^{I} \sum_{j=1}^{J} z_{ij}$.

Un tableau statistique à deux dimensions peut aussi être assimilé à une matrice **aléatoire**.

(iv) En **analyse des données**, on note ainsi $T = (t_{ij})_{(i,j)}$ un tableau statistique initial, où $i \in \{I,..., n\}$ (n observations) et $J \in \{1 ,..., p\}$ (p variables) ; le tableau faisant l'objet de l'analyse directe est souvent obtenu par **transformation des données** $t_{ij} \mapsto x_{ij}$, soit $X = (x_{ij})$.

(v) Dans l'analyse d'une **régression** ou d'un **modèle d'interdépendance**, on partitionne l'ensemble des K + G variables étudiées en K variables dites exogènes et G variables dites endogènes : chacune d'elles est observée N fois, ce qui conduit au tableau statistique suivant, dans lequel les variables, tant endogènes qu'exogènes, peuvent, dans certains types de modèles, être aussi bien quantitatives que qualitatives.

y_{11}	…	y_{1G}	x_{11}	…	x_{1K}
…	…	…	…	…	…
y_{n1}	…	y_{nG}	x_{n1}	…	x_{nK}
…	…	…	…	…	…
y_{N1}	…	y_{NG}	x_{N1}	…	x_{NK}

= [Y , X] = T.