TABLEAU DE CONTINGENCE (C1, D2, F, G, I, J, K1)
(20 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

La notion de **tableau de contingence** peut relever de deux approches :

(a) approche probabiliste : un tableau de contingence est alors une **loi qualitative** associée au croisement des « critères » (ou « codes ») qui le définissent. Ce tableau est **inobservable** puisqu'il s'agit d'une **lp** susceptible d'engendrer des **données** ;

(b) approche statistique (ou « empirique ») : un tableau de contingence synthétise un comptage d'**unités statistiques** classées selon les critères les précédents. Les **effectifs** (à valeurs dans **N**) ainsi dénombrés permettent de calculer des **proportions empiriques** (à valeurs dans **Q**, ou seulement dans **D**). Ce tableau est **observable** puisqu'il contient des données générées par la loi précédente. C'est un **tableau statistique** particulier, dans lequel les « cases » sont quantifiées à l'aide de nombres entiers (effectifs), ou de nombres rationnels déduit des nombres précédents par « normalisation » (division par le total général du tableau) (cf **fréquence absolue**, **fréquence relative**).

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé**, $((\mathcal{X}_h , \mathcal{B}_h))_{h=1,...,H}$ une suite finie d'**espaces mesurables** (ou d'**espaces d'observation**). A chaque espace $(\mathcal{X}_h , \mathcal{B}_h)$ on associe une **partition** (mesurable) finie $\Pi_h = (B_{h,m(h)})_{h,m(h)}$ (ou une **suite** finie de **parties mesurables** disjointes $B_{hm(h)}$ de $\mathcal{X}_h$).

On note $m_h \in \mathcal{T}_h = \{1 ,..., M_h\}$ et $(m_1 ,..., m_H) \in \mathcal{T} = \Pi_{h=1}^{H} \mathcal{T}_h$ (multi-**indice**, ou H-indice) (ou les symboles m(h) et M(h)) désignent, par commodité, resp m_h et M_h). Le H-uple $(m_1 ,..., m_H)$ est souvent appelé « **case** » du tableau.

On considère alors des **variables qualitatives** (ou **« descripteurs »**) $\xi_h : \Omega \mapsto \mathcal{X}_h$ (où h = 1,..., H), qui qualifient les éléments ω de l'**ensemble** fondamental Ω constitué d'**unités statistiques** (individus d'une **population**, le plus souvent).

Par suite, le H-uple aléatoire $\xi = (\xi_1 ,..., \xi_H) : \Omega \mapsto \Pi_{h=1}^{H} \mathcal{X}_h = \mathcal{X}$ ainsi défini correspond à H caractères observés sur chaque élément ω de Ω.

(ii) Soit $\{a_1 ,..., a_N\} = A$ un **échantillon** extrait de Ω et comportant N unités. Le nombre :

(1) $\Pi_{h=1}^{H} \mathbf{1}(\xi^{-1}_h (B_{h,m(h)}))(a_n)$

indique si l'unité a_n appartient, à la fois, à tous les $\xi^{-1}_h (B_{hm(h)})$ (h = 1,..., H) (auquel cas, il vaut 1) ou non (auquel cas, il vaut 0).

Par suite, le nombre entier :

(2) $\Sigma_{n=1}^{N} \Pi_{h=1}^{H} \mathbf{1}(\xi^{-1}_h (B_{h,m(h)}))\ (a_n)$

est le nombre d'unités qui appartiennent à tous les $B_{h,m(h)}$ (h = 1,..., H) : ce nombre est appelé **fréquence (absolue)** de la « **case** » $(m_1, ..., m_H)$ et on le note $n_l = n_{m(1)...m(H)}$, où $(m_1, ..., m_H)$, aussi noté (m(1) ,..., m(H)), est un multi-indice (ou H-indice).

On appelle **tableau de contingence multivarié**, ou **tableau de contingence multidimensionnel** (ou **H-dimensionnel**), **empirique** la suite $T = (n_{m(1)...m(H)})_{m(1)...m(H)}$ ainsi définie : en effet, chaque élément $\omega \in \Omega$ est repéré à partir de H variables ξ_h .

Un tableau (ou table) de contingence est aussi appelé(e) **tableau de dépendance**, ou encore **tableau de(s) correspondance(s)** ou **tableau d'association** (cf **association**, **correspondance**, **dépendance**).

On a donc $\Sigma_{m(1)...m(H)}\ n_{m(1)...m(H)} = N$ (taille de l'échantillon).

(iii) En pratique, on observe directement H variables qualitatives κ_h (h = 1,..., H) sur un échantillon A comportant N éléments (card A = N). Chaque **variable qualitative** κ_h possède M_h « **modalités** » $k_{h,m(h)}$ (h = 1,..., M_h) et l'on peut mettre κ_h en correspondance avec la variable ξ_h selon :

(3) $\kappa_h = k_{h,m(h)} \Leftrightarrow \xi_h \in B_{h,m(h)}$.

Autrement dit, on « discrétise » de façon « finie » chaque variable ξ_h à l'aide d'un **codage** déterminé par chaque variable κ_h (cf **variable quantitative**, **discrétisation**).

Souvent, le nombre entier $n_{m(1)...m(H)}$ reçoit une interprétation concrète en termes de flux entre les parties $B_{h,m(h)}$: dans ce cas, une table de contingence représente aussi un **tableau des transitions**, ou **tableau de transition**, d'**unités statistiques** entre classes (eg **catégories**, **strates**).

Très souvent, H = 2 (tableau à deux dimensions) et les notations se simplifient. Le calcul matriciel intervient alors de façon naturelle puisqu'on peut associer à un tel tableau un **opérateur linéaire** et sa **matrice**.

(iv) On distingue parfois entre :

(a) tableau de contingence à **« marges » fixées** (ou imposées) ;

(b) tableau de contingence à marges arbitraires (ou libres).

(c) tableau à ajuster sur des marges données (cf **ajustement d'un tableau statistique**) ;

Dans ces cas, on admet qu'un modèle probabiliste simple est sous-jacent au tableau T, qualifié alors de **tableau « empirique »**.

On suppose souvent que le **vecteur aléatoire** $\nu : \Omega \mapsto \mathbf{N}^{m(o)}$ (où $m_o = \prod_{h=1}^{H} m_h$ est noté m(o)) défini par les fréquences absolues n_l ($l \in \mathcal{T}$) admet pour **loi de probabilité** la **loi multinômiale** $\mathcal{M}$ (N, p), où $N = \Sigma_{l \in \mathcal{T}} n_l$ et $p_l = p_{m(1)...m(H)} = N^{-1} . n_{m(1)...m(H)}$ est la probabilité (ou fréquence) théorique de la case l, avec $p_l \in [0, 1]$ pour tout $l \in \mathcal{T}$ et $\Sigma_{l \in \mathcal{T}} p_l = 1$ (cf **simplexe**).

Les « probabilités de case » ainsi définies sont généralement estimées à l'aide des fréquences empiriques $f_l = n_l / N$ du tableau T (cf **loi multivariée**, **variable qualitative**).

(v) L'étude d'un tableau de contingence soulève notamment les questions suivantes :

(a) définition d'**indices** (ou de **coefficients**) d'association (ou de dépendance) entre les **variables catégorielles** κ_h (où h = 1,..., H) définie précédemment (cf **coefficient d'association**). Ces indices permettent de définir des **tests d'indépendance** entre critères de la table : **test du chi-deux**, **test du rapport des vraisemblances**, etc ;

(b) ajustement des tableaux sur des marges fixées ;

(c) mise en oeuvre (lorsque H = 2) des méthodes d'**analyse des données** : **analyse des correspondances** ou **analyse sphérique**, etc ;

(d) étude de l'**adéquation** d'un modèle probabiliste (ou d'un **modèle statistique**) sur l'observation synthétique T ;

(e) mise en oeuvre de **modèles à variable dépendante qualitative** (eg **modèle Logit**, **modèle Probit**, etc) ;

(f) étude de la **perte d'information** résultant du passage entre les données élémentaires $\kappa_h (\omega_n)$ (h = 1,..., H ; n = 1,..., N) et les données « agrégées » n_l (resp. f_l) ($l \in \mathcal{T}$) contenues dans T (cf **exhaustivité**, **agrégation**).

(vi) D'un point de vue terminologique, on appelle **tableau de contingence empirique** aussi bien :

(a) le tableau $T = (n_l)_{l \in \mathcal{T}}$ des « niveaux » (cf **fréquence absolue**) ;

(b) que le tableau $F = (f_l)_{l \in \mathcal{T}}$ des **fréquences relatives**, définies par $f_l = n_l / n..$, avec $n.. = N = \Sigma_{l \in \mathcal{T}} n_l$.