ANALYSE DES DONNÉES (K)
(20 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

(i) De façon extensive, la **Statistique** comporte un certain nombre d'activités :

(a) production de **données** d'**observation** : cf **système statistique**, **production statistique**. Ces données concernent des **phénomènes** relevant de chaque **domaine de connaissance** ;

(b) analyse de cette production : cf **Statistique descriptive**, **loi**, **loi scientifique**, **représentation statistique**, **modèle**, **spécification**, **étude scientifique**, **inférence statistique**, **principe de parcimonie** ;

(c) utilisation des résultats et interprétation des analyses : cf **diffusion de l'information**, anticipations ou **prévisions**, **décisions** ou **actions** diverses.

Toutes ces activités concourent donc à analyser des données.

Plus spécifiquement, on désigne par **analyse des données** un ensemble de méthodes, parfois qualifiées de « **méthodes descriptives** », ou « **méthodes sans modèles** », visant :

(a) à fournir une représentation synthétique d'informations, ou observations, souvent nombreuses, généralement multidimensionnelles, regroupées dans de **grandes bases de données**, que l'on peut stocker et traiter numériquement ;

(b) à réduire, ou à « résumer », l'information contenue dans divers **tableaux statistiques**.

Cette approche se caractérise par deux idées a priori :

(a) absence de référence à un **problème statistique** (inférentiel) préalable : « les faits doivent parler d'eux-mêmes » ;

(b) réduction du nombre de variables, ou d'observations, à prendre en compte, détermination d'éventuels **facteurs « cachés »** (cf **variable cachée**), classification ou discrimination des observations, interprétation simple des résultats.

On peut distinguer trois grandes familles de méthodes :

(a) les méthodes d'**analyse factorielle**, qui concernent diverses **variables statistiques** observables sur des **unités statistiques**. Ces variables servent à déterminer des facteurs plus ou moins apparents, dont le rôle est supposé sous-jacent à chaque **phénomène** considéré ;

(b) les méthodes de **classification**, qui concernent des **ensembles** d'**unités statistiques**, à travers diverses **variables** observées (ou mesurées)

sur ces unités. Ces ensembles d'unités sont partitionnés (soit strictement, soit de façon « floue ») en sorte que les classes obtenues soient, à la fois, significatives d'un phénomène et aussi différenciées que possible entre elles : **homogénéité** interne aux classes, **hétérogénéité** externe entre classes ;

(c) les méthodes de **discrimination**, qui ont pour objectif d'imputer une unité statistique donnée à une catégorie parmi d'autres.

(ii) Les méthodes de l'analyse des données se relient naturellement à la **Statistique descriptive**. Comme cette dernière, cependant, les concepts fréquentiels ou descriptifs de l'analyse des données possèdent des correspondances probabilistes. On peut ainsi recenser les suivantes :

(a) **caractère statistique** d'une part, **variable aléatoire** d'autre part;

(b) **histogramme** ou stéréogramme d'une part, **densité** d'autre part ;

(c) **tableau de contingence** d'une part, **loi qualitative** de l'autre ;

(d) caractéristiques « empiriques » (**distances** ou équivalents, **corrélations**, **variabilités**) calculées à partir des données, qui servent souvent de **statistiques** (cf **statistique naturelle**), d'une part, vs **caractéristiques légales** « théoriques » déduites d'une **loi de probabilité**, d'autre part.

Les « données » utilisées elles-mêmes résultent souvent d'un mode d'élaboration spécifiquement statistique (cf **production statistique**) : plans de sondages (ou recensements), plans d'expériences divers.

Cependant, l'analyse des données ne peut se contenter d'une approche formelle (statique ou passive), sans recherche d'une compréhension des phénomènes. Dans ce sens, elle peut aider à la **spécification** d'un **modèle statistique** ou d'un problème de **décision statistique**. En effet :

(a) l'absence d'une théorie relative au phénomène étudié (cf **loi**, **loi scientifique**, **relation fonctionnelle**) peut limiter les avancées de la recherche scientifique, surtout en présence de données (**variables** et **observations**) nombreuses ;

(b) dans ce cas, l'analyse des données peut aider à analyser et déceler des particularités (plus ou moins simples) internes à ces données : variables ou relations pertinentes (mise en évidence de structures et de fonctionnements), ou à forte quantité d'**information** (possibilité de **parcimonie**), etc.

(iii) Font ainsi communément partie de l'analyse des données les branches suivantes de la Statistique :

(a) **Statistique descriptive** ;

(b) **analyse en facteurs communs et spécifiques** (C. BURT, C. SPEARMAN, L.L. THURSTONE) ;

(c) **analyse des correspondances** et ses variantes (J.P. BENZECRI) ;

(d) **analyse en composantes principales** (H. HOTELLING, K. PEARSON) sur des données en niveau ou sur les rangs (**analyse des rangs**) ;

(e) **analyse des covariances partielles** (ou des corrélations partielles) ;

(f) **analyse de la variance** (dans ses aspects descriptifs) ;

(g) **analyse canonique** ;

(h) **régression orthogonale**, **régression sur composantes principales**, **analyse de la variance** ou **analyse de la covariance** (dans leurs aspects descriptifs) ;

(i) méthodes de **classification** ou de classement : notamment, **classification automatique** et **discrimination**.

Moyennant des **transformations des données** (constituées par un tableau initial T), l'**analyse générale des données** synthétise certaines des méthodes précédentes (**analyses factorielles**) en une méthode commune, appliquée au tableau transformé $X = \Psi (T)$: cette méthode commune est basée sur la **décomposition spectrale** de X.

(iv) On considère parfois l'**analyse des données** comme un ensemble de méthodes « descriptives » visant à résoudre un « **problème statistique concret** » sans référence à une **structure statistique** explicite (absence de **modélisation** aléatoire préalable), ni à un **problème statistique** proprement dit.

Cependant, on peut souvent (au moins formellement) « plonger » le formalisme de l'analyse des données dans celui d'un **modèle statistique**, parfois même (eg analyse discriminante ou classification) dans celui de la **décision statistique** (avec adjonction au modèle d'un **espace de décision** et d'une **fonction de perte**).

Ainsi, soit $\xi : \Omega \mapsto \mathbf{R}^K$ un **vecteur aléatoire** réel : eg K **variables** ou **facteurs** mesurés sur N unités statistiques selon la **matrice d'observation** $X \in M_{NK}(\mathbf{R})$. Si ξ est de carré intégrable :

(a) l'**espérance** $E\ \xi = \mu$ admet un **estimateur** tq la **moyenne empirique** (vectorielle) $\bar{X}_N = X'\ e_N / e_N'\ e_N$;

(b) la **dispersion** $V\ \xi = \Sigma$ admet un **estimateur** tq la dispersion empirique (« corrigée » de son biais) $S_N = X'\ P\ X / (N-1)$ (où P désigne la **matrice de centrage** pr à la moyenne empirique).

D'après l'égalité tautologique :

(1) $\xi = E\,\xi + (\xi - E\,\xi) = E\,\xi + \varepsilon = \mu + \varepsilon,$ avec $E\,\varepsilon = 0$ et $V\,\varepsilon = \Sigma,$

l'analyse théorique de ξ revient (eg lorsque $\xi \sim \mathcal{N}_K(\mu, \Sigma)$) à analyser les propriétés de $\Sigma \in S_K(\mathbf{R}) \cap M_K^+(\mathbf{R})$, ce qui se fait habituellement en calculant sa **décomposition spectrale** :

(2) $\Sigma = Q' \Lambda Q = \sum_{\alpha=1}^{rg\,\Sigma} \lambda_\alpha\, q_\alpha\, q_\alpha'$.

L'**analyse générale des données** appliquée à X consiste à décomposer l'équivalent empirique S_N de Σ, ie (cf **statistique naturelle**) :

(3) $S_N = R_N' L_N R_N = \sum_{\alpha=1}^{r} l_\alpha(N)\, r_\alpha(N)\, r_\alpha(N)',$ avec $r = rg\, S_N$.

Un problème statistique « vrai » (dans le cadre aléatoire implicite) peut notamment consister à étudier, à partir de la **loi** P^ξ de ξ, la loi des **valeurs propres** empiriques $l_\alpha(N)$ ($\alpha = 1,..., rg\, S_N$), considérées comme estimateurs naturels des valeurs propres théoriques λ_α , et à tester la significativité (ie la non nullité) de certaines de ces dernières, etc (cf **statistique naturelle**).

(iv) De façon générale, le matériau de base utilisé en analyse des données est constitué de K **variables quantitatives** ξ_1 ,..., ξ_I (formant un vecteur ou une « liste » ξ) et de J **variables qualitatives** (souvent des **variables indicatrices** η_1 ,..., η_J (formant un vecteur ou une liste η). Si $(\Omega, \mathcal{T}, P)$ est un **espace probabilisé** fondamental et si $\mathcal{Z} = (\prod_{i=1}^{I} \mathcal{X}_i) \times (\prod_{j=1}^{J} \mathcal{Y}_j)$ désigne un **ensemble** d'**observation** muni d'une **tribu** adaptée $\mathcal{D}$ (cf **espace d'observation**), on peut considérer que $\zeta = (\xi, \eta) : \Omega \mapsto \mathcal{Z}$ est une **va** de **loi** P^ζ, dont il s'agit d'étudier les propriétés (notamment à travers la **matrice des covariances** ou la **matrice des corrélations**).

(v) En pratique, on dispose d'un N-**échantillon** $Z : \Omega \mapsto \mathcal{Z}^N$ de ζ qui est partitionné de façon analogue, ie Z = (X, Y) (**tableau statistique** à N lignes et K + J colonnes), et l'on cherche à décrire le « comportement » de ζ (ou de P^ζ) à l'aide de Z (ou de la **loi empirique** associée).

On peut, de façon duale, étudier aussi les situations individuelles des N observations (X_n , Y_n) (lignes de la matrice [X, Y]) pr à des groupes ou pr à des moyennes. Lorsque $\mathcal{X}_i = \mathbf{R}$, $\forall$ i, et que, $\forall$ j, η_j (ou $\mathcal{Y}_j$) est codée à l'aide d'une variable à valeurs dans **R** (encore notée, pour simplifier, η_j), l'analyse des données peut mettre en oeuvre les méthodes indiquées ci-dessus (cf **codage**).

(vi) Comme en Statistique, un **tableau synoptique** permet de résumer différents cas (cf aussi **loi multivariée**).

Tableau synoptique

	I variables quantitatives (K exogènes, G endogènes)			J variables qualitatives (L exogènes, H endogènes)		
	G = 0	G = 1	G > 1	H = 0	H = 1	H > 1
K = 0	⌄	Stat. des. quantitative à 1 dim.	Stat. des. quantitative à G dim.	⌄	Stat. des. qualitative à 1 dim.	Stat. des. qualitative à H dim.
K = 1	Stat. des. quantitative à 1 dim.	régression simple		Stat. des. quantitative à 1 dim.	régression simple mixte avec endogène qualitative	
K > 1	Stat. des. quantitative à K dim. (acp, etc)	régression multiple	interdépendance quantitative, analyse canonique, etc	Stat. des. quantitative à K dim.	régression multiple mixte avec endogène qualitative	interdépendance mixte avec endogène qualitative
L = 0	⌄	Stat. des. quantitative à 1 dim.	Stat. des. quantitative à G dim.	⌄	Stat. des. qualitative à 1 dim.	Stat. des. qualitative à K dim. (afc, etc)
L = 1	Stat. des. qualitative à 1 dim.	analyse de la variance à 1 facteur		Stat. des. qualitative à 1 dim.	régression simple qualitative	
L > 1	Stat. des. qualitative à L dim.	analyse de la variance à L facteurs	interdépendance mixte à exogènes qualitatives	Stat. des. qualitative à L dim., afc, etc	régression multiple qualitative	interdépendance qualitative

Notes

Statdes = Statistique descriptive

endogène (resp exogène) = variable endogène (resp variable exogène) ;

dim. = dimension(s) ;

quant. = quantitatif(s), quantitative(s) ; qual. = qualitatif(s), qualitative(s) ;

mixte = existence de variables quantitatives et de variables qualitatives

qualitatif = absence de variables quantitatives

quantitatif = absence de variables qualitatives

Les I variables quantitatives peuvent être constituées de K variables « exogènes » et de G variables « endogènes ». De même, les J variables qualitatives peuvent comporter L **variables exogènes** et H **variables endogènes**.

(vii) Selon la composition des données parmi les variables précédentes, l'analyse peut s'adapter. Ainsi :

(a) la Statistique descriptive (quantitative) peut définir des **histogrammes** ou des **stéréogrammes** à l'aide de G variables numériques (**variables quantitatives**). Cette possibilité peut s'étendre, moyennant convention, à H variables non numériques ;

(b) l'analyse canonique peut porter sur G variables endogènes quantitatives et K variables exogènes quantitatives ;

(c) l'analyse des correspondances peut porter sur un **tableau de contingence** résumant un phénomène décrit à l'aide de H **variables qualitatives** (les « **critères** » de croisement du tableau) ;

(d) l'analyse de la variance ou de la covariance peuvent être concernées par L variables exogènes qualitatives et G variables endogènes quantitatives (analyse multivariée) ; etc.