Glossaire

Algorithme – ensemble des règles opératoires propres à un calcul.

Analyse factorielle – famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à calculer un nombre réduit de "facteurs" résumant approximativement l'ensemble des informations contenues dans le tableau de départ (annexe A1.2).

Analyse en facteurs communs et spécifiques – cette analyse correspond au modèle classique d'analyse factorielle. Elle décrit un ensemble de variables par une combinaison linéaire de facteurs communs sous-jacents et d'une variable (facteur spécifique) synthétisant la part spécifique des variables d'origine (annexe A1.6 ; chapitre 2, section 2.6).

Analyse en composantes principales – méthodes d'analyse factorielle* s'appliquant aux tableaux de mesures (annexe A1.3 ; chapitre 1, section 1.3).

Analyse des correspondances – méthode d'analyse factorielle s'appliquant à l'étude de tableaux à double entrée composés de nombres positifs. L'analyse des correspondances est caractérisée par l'emploi d'une distance (ou métrique) particulière dite distance du chi-2 (ou χ^2) (annexe A1.4 ; chapitre 4, sections 4.3 et 4.4).

Analyse logarithmique – méthode qui consiste à transformer les données en logarithmes (après addition éventuelle d'une constante en cas de données négatives), puis, après les avoir centrées en ligne et en colonne, à les soumettre à une analyse en composantes principales* non normée (annexe A1.5 ; chapitre 5, section 5.7).

Bootstrap – la technique du *bootstrap* consiste à simuler *s* (s généralement supérieur à *30*) échantillons de même taille *n* que l'échantillon initial. Ils sont obtenus par tirage au hasard avec remise parmi les *n* individus observés au départ, ceux-ci ayant tous la même probabilité *1/n* d'être choisis. Cette méthode est employée pour analyser la variabilité de paramètres statistiques simples en produisant des intervalles de confiance* de ces paramètres (annexe A1.9.5 ; chapitre 2, sections 2.3 et 2.4 ; chapitre 4, section 4.4).

Bootstrap partiel – variante du *bootstrap** dans le cas des analyses factorielles qui consiste à projeter les colonnes des tableaux répliqués (mots) comme des éléments supplémentaires* sur les axes de l'analyse de référence, c'est-à-dire l'analyse de l'échantillon inital non perturbé (annexe A1.9.5 ; chapitre 2, section 2.3).

Bootstrap total – variante du *bootstrap* dans le cas des analyses factorielles qui consiste à refaire des analyses en composantes principales* complètes sur chaque échantillon répliqué (annexe A1.9.5 ; chapitre 2, section 2.3).

Bootstrap sur variables – le *bootstrap* est ici réalisé non pas sur les individus, mais sur les variables, ce qui permet d'éprouver les structures observables au niveau des individus (annexe A1.9.5 ; chapitre 2, section 2.4).

Carte auto-organisée de Kohonen – méthode de classification* qui consiste à représenter dans un espace à deux (parfois trois) dimensions un grand nombre de données en respectant la notion de voisinage de l'espace des éléments à classer (annexe A1.8 ; chapitre 3, section 3.1 ; chapitre 4 section 4.2).

Classification – technique statistique permettant de regrouper en classes homogènes des individus ou observations entre lesquels a été définie une distance.

Classification hiérarchique – technique particulière de classification* produisant, par agglomération progressive, des classes ayant la

propriété d'être, pour deux quelconques d'entre-elles, soit disjointes, soit incluses l'une dans l'autre (annexe A1.7 ; chapitre 3, section 3.1).

Coefficient de corrélation – indice exprimant dans quelle mesure deux variables numériques varient de façon concomitante. Cet indice varie de –1 à +1. Il est positif lorsque les valeurs élevées (resp. faibles) d'une variable tendent à être associées aux valeurs élevées (resp. faibles) de l'autre variable. Il est négatif lorsque les valeurs élevées d'une variable tendent à être associées aux valeurs faibles de l'autre variable (annexe A1.3 et *passim*).

Comparaison multiple – un problème de *comparaison multiple* se pose lorsqu'on réitère un test statistique conçu, dans son principe, pour n'être réalisé qu'une seule fois (annexe A1.9.2).

Corpus – ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.

Dendrogramme – représentation graphique d'un arbre de classification hiérarchique*, mettant en évidence l'inclusion progressive des classes (chapitre 3, section 3.1.1).

Distance du chi-2 – distance entre profils* de fréquences utilisée en analyse des correspondances* et dans certains algorithmes* de classification* (annexe A1.4).

Eléments actifs – ensemble des éléments servant de base au calcul des axes factoriels, des valeurs propres* relatives à ces axes et des coordonnées factorielles.

Eléments supplémentaires (ou illustratifs) – ensemble des éléments ne participant pas aux calculs des axes factoriels, pour lesquels on calcule *a posteriori* des coordonnées factorielles (annexe A1.9.3).

Facteur – variables artificielles construites par les techniques d'analyse factorielle permettant de résumer (de décrire brièvement) les éléments actifs* initiaux (variables actives et individus actifs).

Facteur *taille* – Un tel facteur* apparaît lorsque toutes les variables sont corrélées positivement entre elles. Cette caractéristique apparaît le plus souvent sur le premier axe de l'analyse en composantes principales*, que l'on appelle alors "facteur *taille*" (chapitre 5).

Intervalle de confiance – Il permet d'évaluer la précision d'un estimateur et s'interprète comme une marge d'erreur liée au phénomène de la fluctuation d'échantillonnage.

Pourcentages d'inertie (ou de variance) – quantités proportionnelles aux valeurs propres*, dont la somme est égale à 100. Notées τ_α. Dans le cas d'une analyse en composantes principales, les pourcentages d'inertie ne peuvent être systématiquement interprétés en termes de « pourcentages d'information ». Ils peuvent parfois être très faibles et cependant très significatifs statistiquement (cas de données fortement bruitées).

Profil – (d'une ligne ou d'une colonne d'un tableau de contingence*) vecteur constitué par le rapport des effectifs composant une ligne (resp. colonne) à la somme des mêmes effectifs.

Question fermée – question dont les seules réponses possibles sont proposées explicitement à la personne interrogée.

Question ouverte – question posée sans grille de réponse préétablie, dont la réponse peut être numérique (ex: *Quel est votre âge ?*), ou textuelle (exemple, après certaines questions fermées : *Pourquoi ?*) (Chapitre 4).

Tableau de contingence – synonyme de tableau de fréquences ou de tableau croisé : tableau dont les lignes et les colonnes représentent respectivement les modalités de deux questions (ou deux variables nominales), et dont le terme général représente le nombre d'individus correspondant à chaque couple de modalités (annexe A1.4).

Valeurs propres – quantités permettant, lors d'une analyse factorielle*, de juger de l'importance des facteurs* successifs de la décomposition factorielle. La valeur propre notée λ_α mesure la dispersion (variance) des éléments sur l' axe α (annexe A1.3).

Valeur-test – quantité permettant d'apprécier la signification de la position d'un élément supplémentaire* (ou illustratif) sur une axe factoriel. Brièvement, si une valeur-test dépasse 2 en valeur absolue, la position de l'élément correspondant a peu de chance d'être due au hasard (annexe A1.9.1, et *passim)* ; mais attention aux comparaisons multiples* !