INTRODUCTION À LA TROISIÈME PARTIE
L'ANALYSE MULTIVARIÉE : UNE CLASSIFICATION DES MÉTHODES

L'analyse multivariée, au sens large, désigne l'ensemble des méthodes d'analyse statistique qui traitent simultanément plus d'une variable. C'est à l'analyse multivariée que l'on recourt notamment pour

- mesurer le degré d'association entre deux ou plusieurs variables ;
- estimer les paramètres d'une relation entre deux ou plusieurs variables ;
- évaluer à quel point les différences entre deux ou plusieurs groupes d'observations sont significatives ;
- tenter de prédire à quel groupe appartient un individu, à partir de ses autres caractéristiques ;
- essayer de discerner une structure dans un ensemble de données.

Plusieurs techniques d'analyse multivariée distinguent les variables *dépendantes* et les variables *indépendantes*. Les variables dépendantes sont celles dont on veut prédire la valeur ; les autres variables sont appelées indépendantes [1]. On peut classer les méthodes d'analyse multivariée selon le nombre de variables dépendantes et indépendantes, et selon que les unes et les autres sont des variables discrètes ou continues [2].

Le tableau suivant présente un classement de quelques méthodes d'analyse multivariée.

---

1 Pour une discussion des termes « variable dépendante » et « variable indépendante », voir plus loin.

2 Cela découle de l'échelle de mesure associée à chaque variable (voir le chapitre 1-1) : les variables catégoriques sont discrètes, alors que les variables rationnelles et les variables d'intervalle sont traitées le plus souvent comme continues. Pour ce qui est des variables ordinales, il existe peu de méthodes qui leur soient spécifiquement adaptées ; en pratique, on les traite souvent comme continues, mais alors l'interprétation des résultats doit tenir compte de la nature ordinale des variables.

Variable dépendante		**Variables indépendantes**	**Méthode**	
Aucune		2 variables catégoriques	Analyse de tableau de contingence	... à 2 dimensions
		Plus de 2 var. catégo.		... à plus de 2 dimensions
Continue		Discrètes (catégoriques)	Analyse de variance OU Régression multiple	
		Continues et/ou discrètes	Régression multiple	
Catégorique	2 catégories	Continues et/ou discrètes	Logit ou probit	... binomial
	Plus de 2 cat.			... multinomial

Cette partie de l'ouvrage aborde l'analyse de régression. L'analyse de régression est une méthode d'analyse des données qui s'applique lorsque l'on s'appuie sur un modèle théorique formalisé par une relation entre une variable dépendante *continue* et une ou plusieurs variables indépendantes *continues ou discrètes*. Par rapport à la structure fondamentale des données, la régression linéraire adopte le point de vue horizontal (voir 1-1.5). La régression est *linéaire* si la forme fonctionnelle de la relation est linéaire [3].

Avant d'aborder l'analyse de régression comme telle, examinons de plus près la distinction entre les variables *dépendantes* et *indépendantes*, notamment du point de vue de l'examen des liens de causalité. Les termes « variable dépendante » et « variable indépendante » viennent des sciences expérimentales où le chercheur fixe de manière « indépendante » la valeur de certaines variables (comme, par exemple, le dosage d'un traitement), pour observer ensuite l'effet sur la variable « dépendante ». Dans un modèle à une seule équation, la variable dépendante s'appelle aussi « endogène », c'est-à-dire déterminée à l'intérieur du modèle, tandis que les variables indépendantes sont « exogènes », c'est-à-dire déterminées à l'extérieur du modèle. On appelle aussi les variables indépendantes « stimuli » ; les variables dépendantes

sont alors des « réponses ». En anglais, on trouve les couples *predictor/criterion*, *stimulus/response*, *task/performance*, *input/output*.

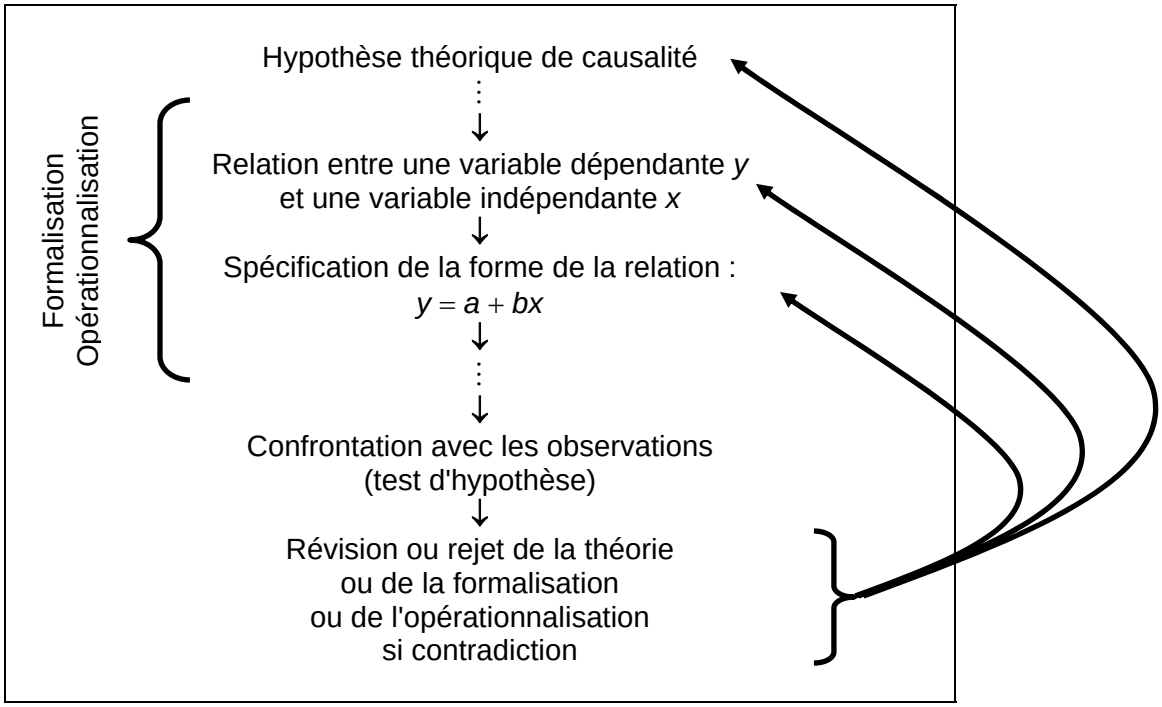
Ce foisonnement de termes est symptomatique de la multitude de significations conceptuelles ou théoriques que peut représenter la relation entre une variable dépendante et une variable indépendante. Les variables indépendantes sont parfois appelées « explicatives ». Cette expression doit cependant s'employer avec prudence, à cause de la connotation de causalité qu'elle véhicule. Il peut arriver que la relation soit purement statistique, de sorte que la variable indépendante peut servir à « prédire » la valeur de la variable dépendante, mais qu'elle n'« explique » pas cette valeur.

Par exemple, on pourrait observer une relation inverse entre la variable dépendante « variation des ventes de détail des boutiques de cadeaux par rapport au mois précédent » et la variable « variation de la température moyenne par rapport au mois précédent ». On aura compris que ce n'est pas le froid qui incite à acheter des cadeaux, mais bien plutôt l'approche de Noël. Il se trouve que dans notre hémisphère, cela coïncide avec l'arrivée de l'hiver; mais en Australie, c'est le contraire. Par contre, il y a bel et bien une relation causale entre la tendance à porter des vêtements plus chauds (variable dépendante) et la tendance au refroidissement saisonnier des températures (variable indépendante). Ce qui permet de faire la distinction entre une relation causale (la seconde) et une relation non causale, c'est le modèle théorique que l'on a du phénomène.

Cela implique notamment que les tests d'hypothèse que l'on pourrait faire sur la relation entre variable dépendante et indépendante ne permettraient pas de démontrer l'existence d'un lien de causalité. Tout au plus pourra-t-on constater que l'hypothèse théorique de causalité est ou n'est pas rejetée par les observations. Ce point est illustré dans le schéma qui suit, que l'on peut mettre en parallèle avec celui de la méthode hypothético-déductive du chapitre 2-2.

---

3 Lorsqu'un modèle théorique ne peut pas se traduire par une relation linéaire, il faut recourir à la régression non linéaire, à laquelle on applique la méthode d'estimation du maximum de vraisemblance.



En somme, la relation entre une variable dépendante et une variable indépendante n'est pas nécessairement une relation causale. L'interprétation que l'on fait de cette relation, la signification qu'on lui donne se rattachent au modèle théorique qui sert de point de départ. Outre les modèles de relation causale, on peut distinguer, par ordre décroissant de contenu théorique :

- des modèles de simulation, parfois appelés modèles de prévision conditionnelle, qui s'appuient sur un modèle représentant le fonctionnement du phénomène [4] et qui sont conçus pour répondre à des questions du type « qu'arriverait-il (ou que serait-il arrivé) si... ? » ;
- des modèles de projection, qui visent à répondre à la question « que va-t-il se passer si la tendance se maintient ? » ou, à partir d'un modèle plus développé du phénomène, « que va-t-il se passer si les paramètres des relations entre les variables demeurent les mêmes ? » ;
- des modèles de prévision, qui ont l'objectif purement pragmatique de répondre à la question « que va-t-il arriver ? » et qui, en fonction de cet objectif, peuvent légitimement exploiter des relations de simple association statistique.

4 Le *comment*, plutôt que le *pourquoi* des modèles de relation causale.