CHAPITRE 11

L'analyse tabulaire multivariée

Au début du chapitre 5, j'ai dit qu'il fallait répondre à six questions lorsqu'on analyse une relation. Nous avons examiné plusieurs façons de répondre aux quatre premières : Y a-t-il une relation dans l'échantillon ? Quelle est l'intensité de cette relation ? Quelle est la direction et la forme de cette relation ? Y a-t-il une relation dans la population ? Pour répondre à ces questions à l'aide d'une analyse tabulaire, nous comparons des pourcentages, calculons et interprétons des mesures d'association, examinons des modèles à l'intérieur des distributions de pourcentages et faisons des tests de signification statistique du chi-carré.

Nous allons maintenant nous attarder aux deux dernières questions : La relation est-elle véritablement causale, ou n'est-elle pas plutôt une relation fallacieuse engendrée par une quelconque tierce variable ? Et, si cette relation est causale, quelle variable intermédiaire relie la variable indépendante à la variable dépendante ? Pour répondre à ces questions, nous devons introduire une ou plusieurs variables supplémentaires dans l'analyse. Dans ce chapitre, nous apprendrons comment répondre à ces questions à l'aide de techniques tabulaires.

Après ce chapitre vous pourrez :

1. Expliquer les conditions qui définissent une relation causale.
2. Reconnaître et expliquer ce qu'est une variable antécédente.
3. Reconnaître et décrire les explications causales et les relations fallacieuses.
4. Reconnaître et décrire les reproductions et les relations véritables.
5. Expliquer ce qu'est une variable dissimulatrice.
6. Expliquer le processus général d'introduction de variables de contrôle (élaboration de tableau).

7. Reconnaître et expliquer ce que sont des variables antécédentes et intermédiaires.
8. Créer et décrire des tableaux multivariés.
9. Calculer et interpréter un gamma partiel.
10. Décrire la relation entre l'analyse multivariée et un devis expérimental.

11.1 La logique des relations causales

Nous sommes tellement habitués à comprendre le monde en termes de causes et d'effets que peu d'entre nous ont déjà réfléchi à ce qu'était vraiment une relation causale. Peu importe la vigueur de nos débats à propos de la causalité dans une association précise, nous tenons pour acquis le concept de causalité. Mais qu'est-ce que cela signifie exactement de dire que la cigarette « cause » le cancer ? Ou que la pauvreté est la « cause » de la criminalité ? Ou que le niveau d'instruction est la « cause » du temps passé devant la télévision ? En termes généraux, qu'est-ce que cela signifie de dire qu'une variable indépendante X cause une variable dépendante Y ?

Nous pouvons affirmer que la variable X est une cause de la variable Y si, et seulement si, trois conditions sont remplies :

1. La variable indépendante X doit « survenir » avant la variable dépendante Y.
2. Les variables X et Y doivent être associées l'une à l'autre.
3. L'association entre les variables X et Y ne doit pas être due à un troisième facteur, une variable antécédente.

Considérons l'une après l'autre ces trois conditions de la causalité.

D'abord, la variable indépendante X doit se produire avant la variable dépendante Y. Autrement dit, la cause doit précéder l'effet. Autrement le monde marcherait à l'envers. C'est pour cela que les films présentés à reculons nous font rire. Des automobiles qui se « détélescopent », des corps qui tombent vers le haut, et de la fumée qui retourne dans le canon d'un fusil, tous ces événements entrent en profonde contradiction avec notre compréhension de l'ordre temporel « convenable » de la relation causale. Des films tel que *Retour vers le futur* nous montrent l'étrangeté d'un monde dans lequel l'ordre temporel des événements est violé.

Bien qu'il soit crucial de déterminer si une relation remplit cette première condition de la causalité, il ne s'agit pas là d'un problème statistique. L'ordonnancement temporel est un problème qui est de

l'ordre de la théorie ou de la méthode de recherche. Notre théorie (au sens le plus large) peut soutenir que les variables sont ordonnées d'une façon particulière ; ainsi nous savons que, en théorie, la variable indépendante se produit avant la variable dépendante. Il semble tenir du simple bon sens que, par exemple, la plupart des adultes aient terminé leur scolarité avant de répondre à des questions concernant le nombre d'heures qu'ils passent devant la télévision. Ou encore, la recherche peut être faite de façon à assurer un ordre correct entre les variables indépendantes et dépendantes. Les expériences sont structurées de telle façon que le chercheur manipule la variable indépendante avant d'observer les effets de cette manipulation sur la variable dépendante. Qu'elle soit résolue par le sens commun, la théorie ou le devis de recherche, la question de l'ordre temporel n'est pas un problème d'analyse statistique des données[1].

Une deuxième condition de la causalité : les variables X et Y doivent être corrélées. Autrement dit, certaines valeurs de la variable dépendante doivent être liées à certaines valeurs de la variable indépendante de façon plus fréquente que ce à quoi nous nous attendrions si seul le hasard jouait. Les taux de cancers doivent être plus élevés chez les fumeurs, le taux de criminalité doit être plus élevé dans les zones les plus pauvres, le nombre d'heures d'écoute de la télévision doit être plus élevé chez les moins instruits avant que nous puissions dire que l'une de ces variables est la cause de l'autre.

Ceci est une préoccupation de la statistique et nous l'avons traité en profondeur déjà. Les techniques bivariées des chapitres 5 à 10 visaient exactement à vérifier cette condition de la causalité. L'analyse tabulaire bivariée, l'analyse de variance, la régression et la corrélation sont des techniques qui servent à déterminer si les variables X et Y sont associées l'une à l'autre. Nous pouvons décider si une variable dépendante est associée à une variable indépendante en comparant les pourcentages dans un tableau bivarié, en comparant les moyennes de la variable dépendante entre les catégories de la variable indépendante, ou en évaluant un diagramme de dispersion et un coefficient de corrélation.

Subsiste une troisième condition : l'association entre les variables X et Y ne doit pas être due à un troisième facteur, une variable antécédente. ***Une variable antécédente*** est une variable qui agit avant la variable indépendante (et par conséquent avant la variable dépendante) dans une chaîne causale. Si c'est une telle variable antécédente qui fait que X et Y sont associés, alors X et Y ne sont pas reliés

1. En fait, quelques techniques statistiques avancées permettent de répondre à cette question dans les situations ambiguës, mais elles dépassent de beaucoup ce que nous pouvons examiner dans ce manuel d'introduction.

de façon causale. Cette troisième condition est, elle aussi, une préoccupation de la statistique. Ce chapitre couvre quelques techniques tabulaires visant à déterminer si cette troisième condition est remplie. Le chapitre suivant nous montrera comment étendre l'analyse de régression et de corrélation au cas où les deux variables sont mesurées sur des échelles d'intervalles ou de proportion.

11.2 Les relations fallacieuses

Voyons comment utiliser l'analyse tabulaire avec des variables antécédentes. Il est souvent plus intéressant d'analyser des données réelles mais il vaut mieux commencer avec un exemple imaginaire. Considérez la relation du tableau 11.1 entre le nombre de cigognes et le taux de natalité dans 200 districts d'un pays européen imaginaire.

Tableau 11.1. Le taux de natalité selon le nombre de cigognes (en pourcentages)

	Nombre de cigognes	
Taux de natalité	Peu	Beaucoup
Élevé	44	62
Bas	56	38
Total	(100)	(100)

$\chi^2 = 6{,}50$; $p < 0{,}05$; $G = 0{,}35$

J'en conviens. Du point de vue du contenu, mon exemple est un peu ridicule. Je l'ai choisi volontairement pour cette raison. Je sais que ce ne sont pas les cigognes qui apportent les bébés. Mais cet exemple fictif me permet d'inventer les fréquences et les pourcentages dont j'ai besoin afin d'illustrer la logique des variables de contrôle. Les exemples hypothétiques ont le merveilleux avantage de présenter des situations moins ambiguës que celles que l'on observe dans le monde réel. Il sera toujours temps de revenir à la réalité plus tard... quand vous travaillerez avec des données réelles.

Alors voyons ce que nous dit le tableau 11.1. Il indique qu'il y a une relation claire entre le nombre de cigognes et le taux de natalité. Seulement 44 % des districts qui ont peu de cigognes ont un taux de natalité élevé, contre 62 % des districts qui ont beaucoup de cigognes, une différence de 18 points de pourcentage. La présence des cigognes est associée à un haut taux de natalité. Le gamma est de

0,35[2]. Un test du chi-carré indique que la relation est significative au seuil 0,05.

Le caractère fantaisiste de cette relation n'empêchera sûrement pas les plus critiques d'entre vous de dire : « Un instant ! Les cigognes n'ont rien à voir avec les bébés. Cette relation n'est pas une relation causale. Ce qui se produit sans doute, c'est qu'il y a beaucoup de cigognes dans les zones rurales et que, dans ces zones, le taux de natalité est élevé. Je parie que si nous tenons compte de caractère rural ou urbain du district, cette relation apparente disparaîtra. »

Bien lancé, jeunes critiques ! Vous avez peut-être raison. Vérifions cela en examinant la relation entre le nombre de cigognes et le taux de natalité séparément dans les districts ruraux et dans les villes. De cette façon, nous garderons constant le caractère rural ou urbain du district dans notre analyse de la relation entre le nombre de cigognes et le taux de natalité. Peut-être trouverons-nous de cette façon un modèle semblable à celui du tableau 11.2.

Tableau 11.2. Le taux de natalité selon le nombre de cigognes, contrôlant l'effet du type de district (en pourcentages)

	District			
	Rural		Urbain	
	Nombre de cigognes		Nombre de cigognes	
Taux de natalité	Peu	Beaucoup	Peu	Beaucoup
Élevé	80	80	20	20
Bas	20	20	80	80
Total	100	100	100	100
(N)	(40)	(70)	(60)	(30)
	$\chi^2 = 0{,}00$; n.s. G = 0		$\chi^2 = 0{,}00$; n.s. G = 0	

Vous aviez raison. Pour les districts ruraux, 80 % de ceux où l'on trouve beaucoup de cigognes et 80 % de ceux où l'on en trouve peu ont un taux de natalité élevé. Aucune différence donc. Dans les districts ruraux, le taux de natalité n'est pas corrélé avec le nombre de cigognes. Et il n'y a pas de relation non plus dans les villes : 20 % des villes ayant peu de cigognes ont un taux de natalité élevé et 20 %

2. Le gamma est équivalent à une autre mesure d'association, le Q, pour les tableaux de 2 par 2. Le Q est toutefois utilisé d'habitude pour les variables nominales, j'utiliserai donc le gamma ici et dans le reste de l'ouvrage.

des villes ayant beaucoup de cigognes ont un taux de natalité élevé. Pas de différence, pas de relation.

Remarquez que, dans chaque moitié du tableau 11.2, on garde constant le type de district (rural ou urbain). Les 110 cas (c'est-à-dire 40 + 70) à gauche sont des districts ruraux et les 90 cas (c'est-à-dire 60 + 30) à droite sont des villes. En fait, le tableau 11.2 est constitué de deux tableaux bivariés, chacun décrivant la relation entre le nombre de cigognes et le taux de natalité pour les districts d'un type particulier. Un tableau bivarié porte sur les zones rurales, l'autre sur les villes. Ce que l'on trouve à l'intérieur de chacune des moitiés de ce tableau ne peut pas être dû au type de district car tous les cas d'une moitié sont du même type, rural ou urbain. Ces tableaux bivariés, lorsqu'ils sont inclus dans un tableau multivarié, s'appellent ***des tableaux partiels*** ou ***des tableaux conditionnels***. J'utiliserai le premier terme, qui est plus commun.

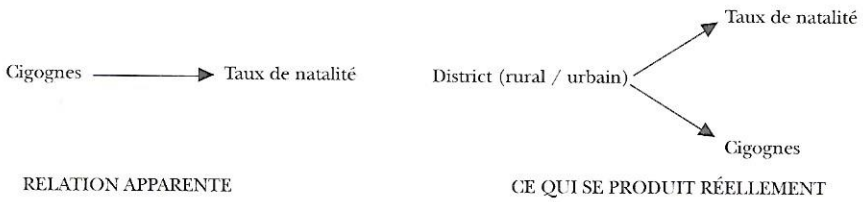
On peut calculer une mesure d'association pour chaque tableau partiel et l'interpréter comme nous le faisions pour les tableaux bivariés. Ici j'ai calculé le gamma (d'autres mesures ordinales – D_{yx} ou tau-b – feraient aussi bien l'affaire, selon la façon dont on veut tenir compte des égalités. Ici je choisis de ne pas tenir compte des égalités). Plus loin dans ce chapitre (à la section 11.10) nous prendrons connaissance d'une mesure d'association, appelée le gamma partiel, qui mesure l'intensité de la relation variable indépendante/variable dépendante, lorsque l'on maintient constante une troisième variable. Pour l'instant cependant, nous nous limiterons aux mesures séparées d'association pour chacun des tableaux partiels.

Nous pouvons aussi faire un test du chi-carré pour le tableau partiel, bien que, dans ce cas, les statistiques deviennent vite relativement complexes et hors de portée de ce livre. Dans le tableau 11.2 j'ai indiqué le résultat du test du chi-carré pour chaque tableau partiel. Ces tests permettent de faire une évaluation convenable de la signification statistique d'une relation partielle. Nous pourrions aussi additionner ces chi-carrés et leurs degrés de liberté en vue de faire un test de signification de la relation entre les variables indépendante et dépendante dans son ensemble, en gardant constante la troisième variable.

Mais il existe des méthodes plus perfectionnées, et franchement meilleures, afin de calculer le chi-carré d'un tableau partiel, des méthodes qui sont basées sur la séparation du chi-carré bivarié en deux chi-carrés différents de tableaux partiels. Ces méthodes sont semblables à l'analyse de variance qui découpe la variance totale en variance inter-groupes et variance intra-groupes. Pour en savoir plus sur l'utilisation du chi-carré pour les tableaux partiels, consultez des manuels plus avancés. Quant à nous, nous nous en tiendrons aux tests « réguliers » du chi-carré pour les tableaux partiels.

Ces tableaux suggèrent que les liens de causalité entre les variables ressemblent à ceux qui sont illustrés dans la figure 11.1. Nous avons vu dans le tableau 11.1 que le nombre de cigognes est lié au taux de natalité – la relation apparente du côté gauche de la figure 11.1. Toutefois, le diagramme causal du côté droit – ce qui se produit réellement – montre que le type de district affecte à la fois les cigognes et les taux de natalité, sans qu'il y ait de relation causale entre les deux.

Figure 11.1. Relation apparente et relation réelle



Nous pouvons vérifier ces relations à l'aide des tableaux 11.3 et 11.4. Eh oui ! Comme nous nous en doutions, les districts ruraux ont plus de cigognes que les villes (tableau 11.3) et les districts ruraux ont un plus haut taux de natalité que les villes (tableau 11.4). Chacune des relations est forte avec un gamma de respectivement – 0,56 et – 0,88 (en traitant le type de district de façon ordinale, avec rural en bas et ville en haut) et statistiquement significative au seuil 0,001.

Tableau 11.3. Le nombre de cigognes selon le district (en pourcentages)

Nombre de cigognes	District	
	Rural	Urbain
Beaucoup	64	33
Peu	36	67
Total	100	100
(N)	(110)	(90)

$\chi^2 = 18{,}18$; $p < 0{,}001$; $G = -0{,}56$

Tableau 11.4. Le taux de natalité selon le district (en pourcentages)

Taux de natalité	District	
	Rural	Urbain
Élevé	80	20
Faible	20	80
Total	100	100
(N)	(110)	(90)

$\chi^2 = 71{,}54$; $p < 0{,}001$; $G = -0{,}88$

Quelques éléments de terminologie

Quelques éléments de terminologie seraient maintenant utiles si l'on veut décrire succinctement la logique de l'analyse multivariée. Une

variable que nous gardons constante pendant l'examen d'une relation bivariée s'appelle ***une variable-contrôle*** ou ***un facteur de test*** (ou facteur test[3]). Le type de district (rural ou urbain) constitue donc une variable-contrôle dans le tableau 11.2. Si nous découvrons qu'une relation bivariée disparaît aussitôt que nous gardons constante une variable antécédente, nous disons que cette relation bivariée est fallacieuse. La relation entre le nombre de cigognes et le taux de natalité décrite au tableau 11.1 est fallacieuse.

Nous utilisons le terme « ***explication*** » pour décrire le résultat de cette analyse. « Explication » est le bon terme car c'est exactement ce que nous faisons : nous *expliquons* la relation bivariée en dégageant la variable antécédente qui en est responsable. Lorsque nous gardons constant le type de district (rural ou urbain), comme nous l'avons fait au tableau 11.2, la relation entre le nombre de cigognes et le taux de natalité disparaît. C'est une relation fallacieuse. Nous l'avons expliquée par la variable antécédente « Type de district » (rural ou urbain).

Le processus général d'introduction de variables de contrôle s'appelle ***l'élaboration.*** Par conséquent, l'explication est un résultat possible de l'élaboration, un résultat qui révèle qu'une relation bivariée est fallacieuse. Nous allons bientôt examiner quelques autres résultats possibles de l'élaboration. Quand des variables de contrôle sont introduites dans un tableau, comme nous le faisons dans ce chapitre, nous appelons ce processus ***élaboration d'un tableau.*** Paul Lazarsfeld, un sociologue réputé qui a enseigné à l'Université Columbia, a formalisé ce que nous appelons le modèle de l'élaboration, et nous lui devons une bonne part de la terminologie que nous employons.

Deux autres termes enfin : nous disons d'une relation bivariée qu'elle est ***d'ordre zéro*** pour la distinguer des relations impliquant une variable de contrôle. Le « zéro » dans l'expression « ordre zéro » signifie qu'il n'y a pas de variable de contrôle (c'est-à-dire zéro variable de contrôle). Par exemple, le tableau 11.1 est un tableau d'ordre zéro, tout comme les tableaux 11.3 et 11.4. Il n'y a pas de variable de contrôle dans ces tableaux, ce sont simplement des tableaux bivariés. Dès que nous introduisons une variable de contrôle, le tableau illustrant chaque catégorie de la variable de contrôle s'appelle ***un tableau partiel.*** Quand nous avons seulement une variable de contrôle, comme au tableau 11.2, nous appelons chacun des « sous-tableaux » ***un tableau partiel d'ordre un.*** La moitié « districts ruraux » et la moitié « villes » du tableau 11.2 sont des tableaux partiels d'ordre un. Si nous introduisons simultanément deux variables de contrôle, nous créons

3. Raymond Boudon utilise l'expression: « variable-test ». Cf. *L'analyse mathématique des faits sociaux*, Paris, Plon, 1970, p. 45 (N.D.T.).

des tableaux partiels d'ordre deux. Ainsi de suite selon le nombre de variables de contrôle que nous introduisons en même temps.

Autrement dit, il y a une explication lorsqu'une relation mise à jour dans un tableau d'ordre zéro disparaît complètement (ou presque complètement) dans des tableaux partiels. On a alors démontré que la relation originale est fallacieuse. Nous avons donc expliqué la relation. Il n'est pas essentiel que la relation disparaisse *entièrement* car cela est peu probable. Nous disons qu'une relation est fallacieuse quand les relations dans les tableaux partiels sont très faibles.

11.4 Des exemples de relations fallacieuses

Les manuels de statistiques et de méthode de recherche présentent des exemples classiques de relations fallacieuses. Je ne peux m'empêcher de vous les présenter. En fait, je les ai déjà mentionnés lorsque je vous ai prévenus, à la section 5.8, que l'association n'implique pas la causalité. Ces exemples sont aussi ridicules que la relation fallacieuse entre le nombre de cigognes et le taux de natalité (qui est un des classiques du genre), mais ils nous aident à comprendre que les relations fallacieuses disparaissent quand une variable-contrôle antécédente est introduite.

Voici le premier exemple. Les inondations du Gange en Inde sont liées au nombre de crimes perpétrés à New York. Par conséquent nous pourrions réduire la criminalité urbaine par le contrôle des crues du Gange ? Aucunement. Les inondations ne sont pas la cause de la criminalité. C'est simplement que la période la plus chaude de l'année provoque des inondations en Inde et augmente la criminalité à New York. La saison est la variable antécédente qui explique la relation inondation-criminalité.

Un autre exemple. Les villes où les prêtres, ministres et rabbins gagnent un meilleur salaire sont celles où se consomment le plus de boissons alcoolisées. Les ministres du culte les mieux payés utiliseraient-ils leurs salaires pour se saouler ? Non. Le salaire de ces personnes et la consommation d'alcool sont deux variables liées à la richesse d'une communauté. Les résidents des villes plus cossues peuvent à la fois mieux payer leurs prêtres, ministres et rabbins, et acheter plus d'alcool. Il n'y a pas de relation causale entre le salaire des ministres du culte et la consommation d'alcool. C'est une liaison fallacieuse.

Encore un autre exemple. Plus il y a de pompiers à combattre un incendie, plus il y a de dommages matériels. Est-ce que les pompiers causent les dommages ? En soufflant sur les flammes, favoriseraient-ils diaboliquement les pertes matérielles ? Eh non ! Une variable

antécédente, l'importance de l'incendie, produit cette association fallacieuse. Plus l'incendie est gros, plus il y a de pompiers. Plus l'incendie est important, plus les dégâts sont imposants. Le nombre de pompiers et la sévérité des dommages ne sont pas liés de façon causale. C'est une association fallacieuse.

Dans la section 5.8 j'ai mentionné qu'il existe une association entre la taille des souliers et les résultats scolaires des élèves du primaire. Je vous laisse le soin de trouver l'explication.

11.5 La reproduction

Rien, dans le processus d'élaboration, ne nous assure qu'une relation bivariée est fallacieuse. Peut-être la relation primitive est-elle « réelle ». Il se peut que la variable indépendante soit la cause de la variable dépendante. Peut-être, mais seulement peut-être, les cigognes apportent-elles les bébés.

Si la relation entre le nombre de cigognes et le taux de natalité décrite au tableau 11.1 est réellement causale, l'introduction d'une variable-contrôle comme le caractère rural/urbain du district devrait produire des tableaux partiels d'ordre un semblables à ceux qui se trouvent dans le tableau 11.5. Chacun de ces tableaux partiels reproduit exactement le tableau bivarié. À la campagne, on trouve des hauts taux de natalité dans seulement 45 % des districts qui ont peu de cigognes mais dans 61 % des districts où il y a beaucoup de cigognes. C'est presque exactement ce que nous avions trouvé dans notre tableau bivarié original. Donc, dans les districts ruraux, le nombre de cigognes est associé au taux de natalité. Je vous épargne le discours analogue pour les villes, mais vous pouvez aisément voir le même modèle d'association entre le nombre de cigognes et le taux de natalité. Les mesures gamma ont sensiblement la même valeur que dans le tableau d'ordre zéro (0,32 et 0,39 en regard de 0,35 au tableau 11.1).

Dans cet exemple, l'introduction de la variable de contrôle (variable-test) « caractère rural du district » ne produit presque aucune différence dans la relation entre les variables indépendante et dépendante. L'association entre le nombre de cigognes et le taux de natalité est à peu près la même dans les districts ruraux que dans les districts urbains. Dans cette situation, nous disons que nous avons procédé à ***une reproduction*** et que nous avons découvert que la relation bivariée primitive était ***véritable.*** (Eh oui ! Encore des termes techniques à retenir !) « Reproduction » et « véritable » sont de bons termes pour décrire cette situation car, dans nos tableaux partiels, nous avons reproduit (c'est-à-dire copié) la relation bivariée primitive et nous avons ainsi découvert qu'elle était véritable.

Tableau 11.5. Le taux de natalité selon le nombre de cigognes, contrôlant l'effet du type de district (en pourcentages)

	District			
	Rural		Urbain	
	Nombre de cigognes		Nombre de cigognes	
Taux de natalité	Peu	Beaucoup	Peu	Beaucoup
Élevé	45	61	43	63
Bas	55	39	57	37
Total	100	100	100	100
(N)	(40)	(70)	(60)	(30)
	$\chi^2 = 2,78$; $p > 0,05$ $G = 0,32$		$\chi^2 = 3,20$; $p > 0,05$ $G = 0,39$	

L'élaboration, donc, peut aboutir soit à une explication, soit à une reproduction (d'autres résultats sont également possibles ; je les décrirai plus loin). La « mécanique » – c'est-à-dire les types de tableaux partiels que nous construisons – est identique pour l'explication et pour la reproduction. Ce qui diffère, c'est ce que les données révèlent. Si les tableaux partiels indiquent qu'il n'y a aucune relation (ou une relation très faible) entre les variables indépendante et dépendante, nous avons une explication : la relation primitive est fallacieuse. Si les tableaux partiels laissent voir sensiblement la même relation entre les variables indépendante et dépendante que celle du tableau primitif, nous sommes en présence d'une reproduction : la relation primitive est véritable.

Bien sûr, dans le cas d'une reproduction, il est possible que la relation primitive soit vraiment fallacieuse et que nous ayons simplement omis d'introduire la bonne variable antécédente. Il subsiste toujours la possibilité que, si nous introduisons une autre variable antécédente, les tableaux partiels révéleront le caractère fallacieux de la relation primitive. Je discuterai de ce problème plus tard lorsque je comparerai le processus d'élaboration à la randomisation, dans les devis expérimentaux.

11.6 Quelque part entre l'explication et la reproduction

L'explication et la reproduction sont deux extrêmes que nous n'atteignons que rarement dans la plupart des analyses réelles. C'est un peu pour cette raison que j'ai choisi un exemple imaginaire. Des données inventées de toutes pièces permettent de construire des

exemples plus clairs que ceux que nous trouvons généralement dans le chaos du monde réel. La plupart du temps, nous découvrons que l'introduction d'une variable antécédente réduit, mais pas complètement, la relation primitive. Le tableau 11.6 présente des tableaux partiels qui illustrent cette situation.

Tableau 11.6. Le taux de natalité selon le nombre de cigognes, contrôlant l'effet du type de district (en pourcentages)

	District			
	Rural		Urbain	
	Nombre de cigognes		Nombre de cigognes	
Taux de natalité	Peu	Beaucoup	Peu	Beaucoup
Élevé	48	57	47	53
Bas	52	43	53	47
Total	100	100	100	100
(N)	(40)	(70)	(60)	(30)
	$\chi^2 = 0{,}95$; p = n.s. G = 0,19		$\chi^2 = 0{,}36$; p = n.s. G = 0,16	

Le nombre de cigognes et le taux de natalité sont toujours liés à l'intérieur de chacun des tableaux partiels du tableau 11.6. En fait les différences en points de pourcentage sont deux fois plus petites que les différences correspondantes du tableau 11.1. Les coefficients gamma sont à peu près deux fois plus faibles (0,19 et 0,16 comparativement à 0,35 dans le tableau d'ordre zéro). Le caractère rural/urbain du district explique en partie seulement la relation entre le nombre de cigognes et le taux de natalité. Cela ne signifie pas que les cigognes sont vraiment la cause des bébés, mais seulement que le caractère rural/urbain du district n'explique pas toute la relation. Il est possible que d'autres variables antécédentes (l'âge ou la dimension des maisons, peut-être) expliquent le reste. Il est possible aussi que ce qui reste « inexpliqué » dans la relation entre le nombre de cigognes et le taux de natalité soit véritable. Une analyse plus poussée, impliquant l'introduction de variables-contrôles additionnelles, peut permettre d'évaluer ces possibilités.

Cette zone grise entre l'explication et la reproduction est beaucoup plus commune que l'explication pure ou la reproduction pure. La raison est que le monde social est composé d'un réseau merveilleusement compliqué et complexe de liaisons multi-causales. Toutes sortes de variables sont reliées à toutes sortes d'autres variables. Étant donné que nous choisissons une variable-contrôle parce que nous croyons que son introduction nous permettra d'obtenir de « bons »

résultats, il est moins probable que nous introduisions une variable-test qui n'ait aucun effet sur la relation bivariée primitive. Par conséquent, la reproduction pure est improbable. De même, le monde social est à ce point multi-causal qu'il est peu probable qu'une variable-contrôle antécédente explique la totalité de l'association entre les deux autres variables. Après tout, peu de variables dépendantes sont liées à une seule variable indépendante qui la précède. Par conséquent, l'explication pure est, elle aussi, bien peu probable.

11.7 La spécification

L'introduction d'une variable-contrôle antécédente peut aussi montrer qu'il y a une relation bivariée pour l'une des valeurs de la variable-test mais pas pour les autres. Prenez par exemple le tableau 11.7. La relation positive d'ordre zéro entre le nombre de cigognes et le taux de natalité se retrouve dans les districts ruraux – en fait elle est un peu plus forte dans les districts ruraux que dans le tableau d'ordre zéro (G = 0,47 comparativement à G = 0,35). Cependant la relation disparaît virtuellement dans les villes (G = –0,03). Nous avons donc *spécifié* où se produit la relation Cigognes → Naissances : dans les districts ruraux mais non dans les villes.

Tableau 11.7. Le taux de natalité selon le nombre de cigognes, contrôlant l'effet du type de district (en pourcentages)

	District			
	Rural		Urbain	
	Nombre de cigognes		Nombre de cigognes	
Taux de natalité	Peu	Beaucoup	Peu	Beaucoup
Élevé	48	71	42	40
Bas	52	29	58	60
Total	100	100	100	100
(N)	(40)	(70)	(60)	(30)
	$\chi^2 = 6{,}23$; $p < 0{,}05$ G = 0,47		$\chi^2 = 0{,}02$; n.s. G = 0,03	

Nous avons de bonnes raisons alors d'appeler ce processus d'élaboration : ***spécification*** (un autre terme que vous devrez connaître). La spécification aussi est un résultat relativement commun du processus d'élaboration. On découvre, par exemple, que plusieurs relations varient pour les hommes et pour les femmes, pour les Noirs et pour les Blancs, pour les gens vivant dans des régions différentes, etc.

11.8 Les variables dissimulatrices

L'explication, la reproduction, quelque chose entre les deux, et la spécification : voici une liste presque exhaustive des résultats qu'il est possible d'obtenir quand nous introduisons une variable de contrôle. Les processus d'élaboration considérés jusqu'à maintenant postulent qu'il y a une relation d'ordre zéro. Notre travail a consisté à trouver une variable antécédente (ou peut-être plus d'une variable) qui explique, affaiblit ou spécifie cette relation primitive.

Mais l'introduction de variables de contrôle s'avère également utile quand les mesures d'association d'ordre zéro entre les variables indépendante et dépendante indiquent une association nulle. Une variable antécédente peut dissimuler une relation qui ainsi n'apparaîtra pas dans un tableau bivarié. Dès que nous introduisons une variable de contrôle appropriée, la « vraie » relation est révélée dans les tableaux partiels. Une telle variable antécédente s'appelle ***une variable dissimulatrice***.

Voici un exemple grotesque. Dans les vrais contes de fée (je ne parle pas des versions épurées de Walt Disney) beaucoup d'enfants sont dévorés par des sorcières et des trolls. Vous êtes-vous déjà demandé si ces pratiques alimentaires donnent des indigestions à ces macabres personnages de contes de fée ? Pourtant il me semble que vous auriez dû ! J'imagine sans mal des sorcières autour de leur marmite, ou des trolls sous un pont, éructer et se plaindre de maux d'estomac après avoir dévoré plusieurs enfants particulièrement dodus qui ont été désobéissants (ou bien étaient-ce des étudiants qui n'avaient pas fait leurs exercices de statistiques ?).

Vérifions la relation entre le fait de manger des enfants et l'indigestion. Le tableau 11.8 est un tableau d'ordre zéro illustrant la relation entre les préférences alimentaires et la fréquence des indigestions chez 500 goules de contes de fée (les données sont fictives !). Eh non ! Il n'y a pas de relation entre les préférences alimentaires et la fréquence des indigestions. Les goules qui n'apprécient pas les enfants sont aussi susceptibles de maux d'estomac que les goules qui préfèrent manger des enfants. Je suppose que le fait de manger des enfants ne cause pas de maux d'estomac chez les goules.

Mais, un instant ! Je parie que les sorcières, plus que les trolls, ont tendance à aimer manger des enfants (rappelez-vous Hansel et Gretel). Et je parie aussi que les sorcières sont plus sensibles de l'estomac que les trolls. Si j'ai raison, le type de goule (sorcière ou troll) est une variable qui dissimule la vraie relation entre la préférence alimentaire et la fréquence des indigestions.

Tableau 11.8. La fréquence des indigestions selon les préférences alimentaires (en pourcentages)

Fréquence des indigestions	Préférences alimentaires	
	Aime bien les enfants	N'aime pas les enfants
Élevé	50	50
Bas	50	50
Total	100	100
(N)	(200)	(300)

$\chi 2 = 0{,}00$; n.s. ; Lambda = 0,00

Le tableau 11.9 est un tableau partiel qui vérifie ce pressentiment. Mais oui, il y a une relation entre les préférences alimentaires et la fréquence des indigestions, bien que cette relation ne soit pas dans la même direction pour les trolls que pour les sorcières. Chez les trolls, le goût pour les enfants est associé à l'indigestion : 80 % des trolls qui savourent les enfants souffrent souvent d'indigestion, contre 30 % seulement des trolls qui n'aiment pas manger les enfants. On trouve le modèle opposé chez les sorcières, où les indigestions tourmentent seulement 20 % de celles qui aiment manger des enfants mais un énorme pourcentage (90 %) de celles qui ne mangent pas d'enfants. Ce qui semblait être une association nulle dans un tableau d'ordre zéro se révèle être une relation complexe quand une variable de contrôle est introduite dans l'analyse. Le type de goule dans cet exemple joue le rôle de variable dissimulatrice qui masque la relation complexe entre les préférences alimentaires et la fréquence des indigestions.

Tableau 11.9. La fréquence des indigestions selon les préférences alimentaires, contrôlant le type de goules (en pourcentages)

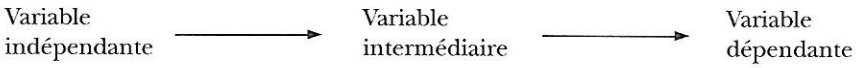
Fréquence des indigestions	Type de goules			
	Troll		Sorcière	
	Préférences alimentaires		Préférences alimentaires	
	Aime bien les enfants	N'aime pas les enfants	Aime bien les enfants	N'aime pas les enfants
Élevée	80	30	20	90
Basse	20	70	80	10
Total	100	100	100	100
(N)	(100)	(200)	(100)	(100)
	$\chi 2 = 66{,}96$; $p < 0{,}001$ Lambda = 0,43		$\chi 2 = 98{,}99$; $p < 0{,}001$ Lambda = –0,67	

Incidemment, bien que je ne le fasse pas ici, vous pouvez construire deux tableaux d'ordre zéro supplémentaires à partir des fréquences marginales du tableau 11.9 : les préférences alimentaires selon le type de goule et la fréquence des indigestions selon le type de goule. Vous verrez que ces tableaux bivariés montrent que les sorcières, plus que les trolls, ont tendance à préférer manger des enfants et à souffrir d'indigestion.

La morale de cette section : si aucune relation n'apparaît dans un tableau d'ordre zéro alors que vous vous attendiez à en trouver une, songez à la possibilité qu'une variable dissimulatrice soit à l'œuvre.

11.9 Les variables intermédiaires comme variables de contrôle

Jusqu'à maintenant nous avons considéré, comme variable de contrôle, les seules variables antécédentes. Mais nous pouvons utiliser le même procédé pour contrôler l'effet d'***une variable intermédiaire*** que nous considérons comme le lien causal entre une variable indépendante et une variable dépendante. La deuxième condition de la causalité, la condition temporelle, exige que la variable intermédiaire apparaisse après la variable indépendante et avant la variable dépendante dans la chaîne causale (voir la section 11.1). Schématiquement, une chaîne causale incluant une variable intermédiaire ressemble à ceci :



Si nous contrôlons l'effet d'une variable qui, croyons-nous, relie une variable indépendante à une variable dépendante, trois choses peuvent survenir. La relation primitive peut disparaître dans les tableaux partiels, la relation primitive peut se maintenir dans les tableaux partiels, ou quelque chose entre les deux situations peut se produire. Considérons ces situations l'une après l'autre.

D'abord si la relation primitive disparaît à l'intérieur des tableaux partiels dans lesquels on contrôle l'effet d'une variable qui semble intervenir entre la variable indépendante et la variable dépendante, nous concluons que la variable intermédiaire relie vraiment les variables indépendante et dépendante. Le modèle de résultats dans les tableaux partiels ressemblerait alors à ceux que nous retrouvons pour les relations fallacieuses, bien que ce soit pour des raisons différentes. Et tout comme nous avons conclu que, dans une relation fallacieuse, la variable antécédente explique toute la relation entre les

variables indépendante et dépendante, nous concluons maintenant que la variable intermédiaire explique entièrement la façon dont la variable indépendante affecte la variable dépendante.

Ce modèle ne signifie pas, bien sûr, que la variable indépendante et la variable dépendante ne sont pas liées. Au contraire, elles le sont certainement. Nous avons même démontré comment la variable indépendante cause la variable dépendante. Cela se produit par l'intermédiaire de la variable intermédiaire. Le nom technique de cette situation est ***l'interprétation.*** Nous disons que la variable intermédiaire ***interprète*** la relation entre la variable indépendante et la variable dépendante.

Mais que se passe-t-il lorsque les tableaux partiels ressemblent aux tableaux primitifs ? Nous concluons alors que la variable que nous croyions être une variable intermédiaire n'intervient pas vraiment. Elle ne lie pas les variables indépendante et dépendante. La variable de contrôle ne fait pas partie de la chaîne causale. Ceci ne veut pas dire qu'il n'y a pas de variable intermédiaire, mais plutôt, plus modestement, que nous n'en avons pas discerné.

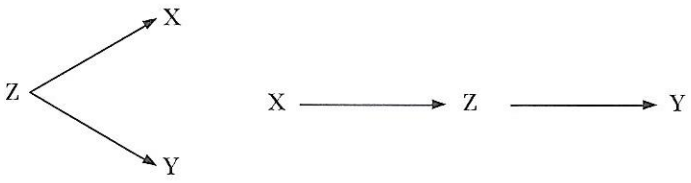
Finalement, la troisième possibilité. À l'instar des variables antécédentes, le contrôle de l'effet d'une variable intermédiaire aboutit souvent, en réalité, quelque part entre ces deux extrêmes où, d'une part, on interprète complètement une relation causale et où, d'autre part, on ne l'interprète pas du tout. Souvent, on découvre qu'une variable intermédiaire plausible interprète une partie seulement de la relation entre une variable indépendante et une variable dépendante. La variable intermédiaire est importante, mais il doit y avoir d'autres variables qui elles aussi relient de façon causale les variables indépendante et dépendante. Eh oui ! Le monde est un réseau touffu de relations causales.

Si nous trouvons une variable qui relie les variables indépendante et dépendante, notre travail n'est pas terminé pour autant. Nous pouvons continuer à chercher des variables additionnelles qui interprètent les liens entre la variable indépendante et la variable intermédiaire et entre la variable intermédiaire et la variable dépendante. Et ainsi de suite, à mesure que nous rendons explicite le réseau de causalité[4]. Je serai franc, il y a des moyens beaucoup plus

4. Les chaînes de causalité soulèvent de fascinantes questions philosophiques et pratiques au fur et à mesure qu'elles s'allongent. Est-ce que les battements d'ailes d'un papillon en Chine ont un effet sur le temps qu'il va faire au Michigan ? Les chaînes causales n'ont-elles pas de fin ou finissent-elles par s'éteindre petit à petit? À partir de quel moment, s'il y en a un, pouvons-nous dire qu'un événement est trop distant d'un autre, trop éloigné dans la chaîne de causalité, pour en être la cause ? La plupart des débats à propos des répercussions des faits historiques sur les problèmes sociaux contemporains portent sur de telles questions.

efficaces que l'analyse tabulaire pour examiner les relations causales. Cependant, la logique qui sous-tend ces méthodes plus avancées est très semblable à la logique de l'analyse de tableau.

Je veux attirer votre attention sur l'importance des modèles théoriques dans l'analyse statistique. La logique et les procédures d'élaboration ne font aucune distinction entre les deux modèles suivants :



Modèle A Modèle B

L'analyse de donnée mettant en relation X et Y tout en contrôlant l'effet de Z donne le même résultat, que ce soit le modèle A ou le modèle B qui décrive la véritable relation entre les variables. L'analyse de données ne nous permettra donc pas de choisir entre les modèles A et B. Ce choix doit être basé sur nos paradigmes, nos théories et même parfois sur notre bon sens.

11.10 Le gamma partiel

Nous avons vu que le gamma est une mesure d'association qui convient aux données d'un tableau bivarié impliquant des variables ordinales ou des variables d'intervalle et de proportion regroupées en catégories. Le ***gamma partiel*** (représenté par le symbole G_p) est l'extension du gamma aux tableaux multivariés. Le gamma partiel décrit l'intensité de la relation entre une variable indépendante et une variable dépendante mesurées au niveau ordinal, une fois l'effet d'une variable-contrôle éliminé.

Le calcul du gamma partiel est une généralisation directe du gamma. Rappelez-vous la formule du gamma donnée à la section 7.5 :

$$G = \frac{\text{Semblables} - \text{Opposées}}{\text{Semblables} + \text{Opposées}}$$

lorsque « Semblables » et « Opposées » renvoyaient au nombre de paires de cas ordonnés dans la même direction ou dans la direction opposée pour les variables indépendante et dépendante. Pour obtenir le gamma partiel, nous comptons les paires semblables et oppo-

sées dans chaque tableau partiel en suivant exactement la même procédure que nous avons suivie lorsque nous travaillions avec des tableaux bivariés (voir la section 7.5). Nous additionnons ensuite le nombre de paires semblables à l'intérieur de chaque tableau partiel ; de même, nous additionnons toutes les paires opposées à l'intérieur de chaque tableau partiel. Nous calculons enfin le gamma partiel à l'aide de cette formule :

$$G_p = \frac{\Sigma\text{Semblables} - \Sigma\text{Opposées}}{\Sigma\text{Semblables} + \Sigma\text{Opposées}}$$

Le signe de sommation Σ commande ici d'additionner les paires pour tous les tableaux partiels.

Par exemple, essayons de trouver le gamma partiel pour la relation partielle d'ordre un décrite dans le tableau 11.6. Ce tableau décrit la relation entre le taux de natalité et le nombre de cigognes, en maintenant constant le type de district. Voici les fréquences sur lesquelles sont basés les pourcentages dans le tableau 11.6 :

	District			
	Rural		Urbain	
	Nombre de cigognes		Nombre de cigognes	
Taux de natalité	Peu	Beaucoup	Peu	Beaucoup
Élevé	19	40	28	16
Bas	21	30	32	14

On calcule le gamma en trouvant d'abord le nombre de paires de cas dont les scores des variables dépendante et indépendante sont ordonnés dans la même direction puis dans la direction opposée :

$$\begin{aligned}\text{Semblables} &= (21)(40) + (32)(16)\\ &= 840 + 512\\ &= 1352\end{aligned}$$

$$\begin{aligned}\text{Opposées} &= (19)(30) + (28)(14)\\ &= 570 + 392\\ &= 962\end{aligned}$$

Donc :

$$G_p = \frac{\Sigma\text{Semblables} - \Sigma\text{Opposées}}{\Sigma\text{Semblables} + \Sigma\text{Opposées}}$$

$$= \frac{1352 - 962}{1352 + 962}$$

$$= \frac{390}{2314}$$

$$= 0{,}168$$

$$= 0{,}17$$

Ce gamma partiel de 0,17 indique une relation relativement faible entre le taux de natalité et le nombre de cigognes, même après avoir éliminé l'effet du type de district. Considérant des paires de districts, nous avons réduit nos erreurs de 17 % en « prédisant » que, dans une paire, le district ayant le plus grand nombre de cigognes a aussi le plus haut taux de natalité, une fois qu'on a éliminé l'effet du caractère rural ou urbain du district. Le gamma partiel est une moyenne pondérée de ces gammas d'ordre zéro. Veuillez noter qu'on peut utiliser les gammas partiels même si la variable de contrôle est nominale pour autant que les variables dépendantes et indépendantes soient au moins de niveau ordinal.

11.11 Résumé de l'élaboration

Il serait maintenant utile de faire le point sur ce qui peut se produire lorsque nous entreprenons une analyse tabulaire multivariée de la relation entre une variable indépendante (VI) et une variable dépendante (VD), en contrôlant l'effet d'une troisième variable (VC). Le tableau 11.10 présente un résumé des situations que nous pouvons rencontrer dans le processus d'élaboration.

Tableau 11.10. Résumé de l'élaboration de tableau

Variable-contrôle VC	Relation VI-VD primitive ?	Relation VI-VD contrôlant l'effet de VC (tableaux partiels)	Conlusion à tirer
Antécédente	Oui	Pas de relation $G_p \approx 0$	La relation VI-VD est fallacieuse. (Explication)
Antécédente	Oui	Semblable à la relation primitive VI-VD $G_p \approx G$	La relation est véritable. (Reproduction)
Antécédente	Oui	Plus faible que la relation VI-VD primitive $0 < G_p < G$	VC explique en partie la relation VI-VD.
Antécédente	Oui	G_p varie d'un tableau à l'autre	La relation VI-VD est tributaire de la valeur de VC. (Spécification)
Antécédente	Non	Il y a relation dans les deux tableaux partiels $G_p > G \approx 0$	VC est une variable dissimulatrice qui masque la relation VI-VD.
Intermédiaire	Oui	Pas de relation $G_p \approx 0$	VC relie de façon causale VI à VD. (Interprétation)
Intermédiaire	Oui	Semblable à la relation primitive $G_p \approx G$	VC ne relie pas VI à VD
Intermédiaire	Oui	Plus faible que la relation primitive $0 < G_p < G$	VC relie en partie VI et VD, mais il existe d'autres variables intermédiaires

11.12 L'élaboration et le problème des petits N

L'élaboration de tableau est peu efficace car elle demande un grand nombre de cas. Plus il y a de variables de contrôle et plus il y a de valeurs dans la variable indépendante et dans la variable de contrôle, plus il faut de cas. Si on n'a pas suffisamment de cas, les pourcentages des tableaux partiels seront basés sur trop peu de cas pour qu'on puisse avoir confiance en nos résultats. Les fréquences attendues peuvent également être trop petites (moins que 5) pour effectuer

des tests de chi-carré. Même dans la situation peu probable où les cas seraient répartis également entre les tableaux partiels, des pourcentages basés sur 100 cas dans des relations d'ordre zéro seraient basés sur 50 cas dans les tableaux partiels contrôlant l'effet d'une variable dichotomique. Et si une seconde variable dichotomique est introduite, ces 50 cas seront réduits à 25 cas pour établir les pourcentages, et ainsi de suite. Le problème est encore plus grave dans le cas des variables de contrôle non dichotomiques (les variables de contrôle qui ont plus de deux catégories). Oui, à l'exception des très grands N, l'élaboration de tableau réduit très rapidement le nombre de cas.

On peut remédier (mais sans les éliminer tout à fait) aux problèmes de la diminution du nombre de cas en réduisant le nombre de catégories de la variable indépendante et de la variable de contrôle. On peut aussi exclure les catégories de la variable indépendante ou de la variable de contrôle qui ont un faible nombre de cas. Nous avons déjà évoqué ces stratégies de gestion des données dans la section sur les tableaux bivariés, et elles fonctionnent aussi bien pour l'analyse multivariée. Comme dans les analyses bivariées, bien sûr, on ne devrait pas agréger ou exclure des catégories si, en procédant ainsi, on élimine des détails importants ou si l'on perd toute la signification théorique.

Toutefois il n'y a pas de technique de gestion de données qui puisse vraiment résoudre les problèmes posés par l'inefficacité de l'élaboration de tableau. En pratique nous en sommes réduits à n'introduire qu'une ou deux variables de contrôle à moins que notre N soit très important. Malgré ces restrictions, l'élaboration de tableau offre un moyen utile, et quelquefois essentiel de tenir compte des variables antécédentes et intermédiaires.

11.13 La relation entre l'analyse multivariée et le devis expérimental

Dans de véritables expériences, le chercheur détermine aléatoirement (c'est-à-dire au hasard) quels sujets feront partie du groupe expérimental et lesquels feront partie du groupe-contrôle. Dans les limites de la randomisation, le groupe expérimental et le groupe-contrôle sont identiques (ou presque) en ce qui concerne chaque variable antécédente possible : sexe, race, culture, etc., incluant même des variables aussi bizarres que le personnage de bande dessinée favori ou la préférence pour les légumes cuits ou crus (qui n'ont sans doute pas de rapport avec le sujet de la recherche). Le groupe expérimental et le groupe-contrôle sont identiques (ou presque) même pour des

variables auxquelles vous ne penserez jamais. (Désolé, mais je ne peux pas vous donner d'exemple de variables auxquelles vous ne penserez jamais.) Les groupes expérimental et de contrôle sont alors équivalents (ou presque) quant à toutes les variables qui pourraient influencer la relation entre la variable indépendante et la variable dépendante. Le hasard « fonctionne » mieux si l'on a des grands nombres ; donc plus on a de sujets dans une expérience, plus le chercheur a de chances que les groupes de contrôle et expérimental soient parfaitement équivalents.

C'est un avantage considérable de l'expérimentation : le devis lui-même impose implicitement des contrôles à toutes les variables antécédentes possibles. Une fois que la variable indépendante est introduite, les différences que le chercheur observe dans la variable dépendante entre le groupe expérimental et le groupe-contrôle ne peuvent être dues à une variable antécédente, quelle que soit cette différence. À l'intérieur des limites de la randomisation, le groupe expérimental et le groupe-contrôle ont été, en quelque sorte, « nivelés » pour toutes les variables antécédentes.

Mais il arrive parfois que nous ne puissions pas faire d'expérimentation. Pour des raisons pratiques ou éthiques, il se peut que nous ne puissions pas assigner aléatoirement les sujets au groupe expérimental et au groupe-contrôle. Si nous étudions les effets du sexe, par exemple, nous ne pouvons pas décider par randomisation quels sujets seront hommes et quels sujets seront femmes. (Y a-t-il des volontaires pour cette expérience ?) Si nous étudions les effets du niveau d'instruction, nous ne pouvons pas décider aléatoirement quels sujets iront à l'université et lesquels arrêteront leur scolarité au secondaire (des volontaires de nouveau ?). Au lieu de cela, nous devons prendre les gens comme ils sont et faire de notre mieux à l'aide de devis de recherche non expérimentaux, tels les sondages.

Et c'est ici qu'entrent en scène les techniques multivariées. L'utilisation de variables-contrôles remplace la randomisation des devis expérimentaux. L'introduction d'une variable-contrôle antécédente élimine l'effet de cette variable sur la relation entre la variable indépendante et la variable dépendante. C'est exactement ce que la randomisation fait, bien qu'elle le fasse plus efficacement en éliminant simultanément l'effet de toutes les variables antécédentes. L'analyse multivariée procède beaucoup plus lentement, en contrôlant l'effet des variables antécédentes à tour de rôle ou, au mieux, quelques-unes à la fois.

Cela signifie que, en ce qui concerne la nature véritable de la relation, nous ne pourrons jamais avoir autant confiance en une analyse multivariée qu'en un devis expérimental. Même si une relation

« tient » après que nous ayons contrôlé l'effet d'une variable antécédente, il est toujours possible qu'une autre variable antécédente non contrôlée explique la relation. Nous pourrions alors introduire une deuxième variable-contrôle, ou peut-être une troisième, ou même plus. Mais les banques de données ont des limites qui font que nous atteignons tôt ou tard le point où il n'y a plus de variables antécédentes à contrôler. D'ailleurs, il se pourrait que nous ne sachions même pas quelle est la variable antécédente appropriée.

Mais nous faisons de notre mieux. Inutile de pleurnicher, il y a trop à faire en recherche. À l'aide des théories et d'autres idées qui éclairent notre travail, nous essayons d'inclure autant de variables antécédentes plausibles qu'il est possible dans notre collecte de données. Nous contrôlons ensuite l'effet de ces variables dans nos analyses. Bien sûr, à mesure que nous introduisons des variables de contrôle, nous pouvons avoir une plus grande confiance en la solidité de la relation. Mais nous devons accepter cette impossibilité de contrôler les effets de chaque variable antécédente possible à moins d'utiliser un devis expérimental. Mais cela dit, nous faisons de notre mieux et, si nos théories, notre devis de recherche et nos analyses statistiques sont solides, ce que nous aurons fait de mieux ne sera pas si mal.

11.14 Résumé du chapitre 11

Voici ce que nous avons appris dans ce chapitre :

- Une variable antécédente intervient avant les variables indépendante et dépendante dans la chaîne causale.
- L'élaboration est le processus qui consiste à analyser une relation bivariée après avoir éliminé les effets d'une ou de plusieurs variables-contrôles. Lorsque l'on se sert de tableaux de pourcentages, ce processus s'appelle l'élaboration d'un tableau.
- Une relation partielle élimine les effets d'une troisième variable, antécédente ou intermédiaire.
- Quand une relation bivariée disparaît dans les tableaux partiels qui contrôlent l'effet d'une variable antécédente, la relation est fallacieuse ; elle est expliquée par la variable antécédente. Ce processus s'appelle l'explication.
- Quand une relation bivariée demeure dans les tableaux partiels qui contrôlent l'effet d'une variable antécédente, la relation bivariée est véritable (en attendant l'introduction d'une autre variable-contrôle antécédente). Ce processus s'appelle la reproduction.

- Il arrive souvent que l'introduction d'une variable-contrôle réduise, sans complètement l'éliminer, la relation bivariée primitive. Ceci indique que la variable-contrôle explique en partie seulement la relation.
- Quand les tableaux partiels sont différents l'un de l'autre, la variable-contrôle « spécifie » la relation bivariée. Ce processus s'appelle la spécification.
- Une variable intermédiaire intervient, dans la chaîne causale, entre la variable indépendante et la variable dépendante.
- Si l'introduction d'une variable-contrôle intermédiaire élimine la plus grande partie d'une relation, la variable-contrôle relie de façon causale la variable indépendante et la variable dépendante. Ce processus s'appelle l'interprétation.
- Si l'introduction d'une variable-contrôle intermédiaire réduit une relation sans l'éliminer, la variable-contrôle relie de façon causale la variable indépendante et la variable dépendante. Il peut toutefois exister également d'autres variables intermédiaires.
- Le gamma partiel est une mesure RPE d'association qui résume la relation entre des variables ordinales, d'intervalles ou de proportion dont les valeurs ont été regroupées en catégories, en contrôlant les effets d'une troisième variable.
- Le contrôle des variables antécédentes dans l'analyse multivariée est une approximation limitée de la randomisation dans la recherche expérimentale.

Principaux concepts et procédures

Termes et idées

analyse tabulaire multivariée
tableau partiel
relation causale
variable antécédente
variable-contrôle ou facteur de test
relation fallacieuse
explication
élaboration et modèle d'élaboration
élaboration de tableau
relation d'ordre zéro

tableau partiel d'ordre un
reproduction
relation véritable
spécification
variable dissimulatrice
variable intermédiaire
interprétation
gamma partiel

Symboles

G_p

Formules

$$G_p = \frac{\Sigma\text{Semblables} - \Sigma\text{Opposées}}{\Sigma\text{Semblables} + \Sigma\text{Opposées}}$$

RAPPORT D'ANALYSE N°7
ANALYSE TABULAIRE MULTIVARIÉE

Nous avons vu précédemment *(dans le rapport n° 3)* que les démocrates et les indépendants sont plus favorables à la discrimination positive envers les femmes que les républicains. Le tableau 1 présente cette relation. La différence entre les républicains et les démocrates quant au soutien de la discrimination positive peut être due au fait que les femmes soutiennent plutôt les démocrates alors que les hommes sont plutôt républicains.

TABLEAU 1 ICI

Le tableau 2 introduit un contrôle selon le sexe. La relation illustrée dans le tableau original est reproduite dans les tableaux partiels. Que ce soit pour les femmes ou les hommes, on retrouve la même relation où les démocrates et les indépendants soutiennent plus fortement que les républicains la discrimination positive envers les femmes. La relation est un peu plus forte pour les hommes que pour les femmes (respectivement un gamma de 0,34 et 0,27). Donc la différence entre les sexes quant aux préférences politiques n'explique pas la différence entre les partis quant au soutien de la discrimination positive envers les femmes.

TABLEAU 2 ICI

Inclure les tableaux à la fin du rapport

Tableau 1. Discrimination positive envers les femmes selon la préférence quant au parti politique (en pourcentages)

Discrimination positive envers les femmes	Préférence quant au parti politique		
	Démocrate	Indépendant	Républicain
En accord	64,2	61,4	42,0
Ni l'un ni l'autre	12,7	15,5	15,1
En désaccord	23,1	23,2	42,9
Total	100,0	100,1	100,0
(N)	(685)	(207)	(517)

$\chi^2 = 71{,}678$; dl = 4 ; $p < 0{,}001$; G = 0,32

Tableau 2. Discrimination positive envers les femmes selon la préférence quant au parti politique (en pourcentages)

Discrimination positive envers les femmes	Sexe					
	Hommes			Femmes		
	Préférence politique			Préférence politique		
	Dém.	Ind.	Rép.	Dém.	Ind.	Rép.
En accord	60,2	57,0	36,3	66,7	64,9	48,0
Ni l'un ni l'autre	12,7	16,1	15,4	12,7	14,9	14,8
En désaccord	27,0	26,9	48,3	20,6	20,2	37,2
Total	99,9	100,0	100,0	100,0	100,0	100,0
(N)	(259)	(93)	(267)	(426)	(114)	(250)
$G_p = 0,30$	$\chi^2 = 36,398$; dl = 4 ; $p < 0,001$; G = 0,34			$\chi^2 = 25,812$; dl = 4 ; $p < 0,001$; G = 0,27		