2 |

CONCEVOIR ET PRÉPARER LES VARIABLES NÉCESSAIRES À L'ANALYSE

1. QUESTIONS, VARIABLES ET MODALITÉS

Il est indispensable de distinguer deux niveaux d'information, même s'ils se recoupent largement. Le premier niveau est celui des réponses fournies par les enquêtés aux questions ou les codages de matériaux qualitatifs. Il s'agit d'informations primaires, liées à la logique et aux exigences spécifiques de l'enquête ou du codage : les informations sont réduites, limitées à un aspect précis mais ponctuel. Il s'agit par exemple des titres des livres lus durant le dernier mois, du nombre d'enfants, de la situation professionnelle, de la destination des dernières vacances, de la durée de la dernière période de chômage...

Ce premier niveau d'informations regroupe les variables qualifiées de « primaires » (au sens de « premières ») et qui sont évidemment indispensables : elles constituent le matériau empirique. Pourtant, ces variables ne sont pas suffisantes ou satisfaisantes dans bien des cas : elles répondent à des exigences empiriques et aux impératifs méthodologiques de réalisation de l'enquête ; elles ne renseignent que partiellement le sociologue sur ce qui l'intéresse en priorité.

Le second niveau est celui des variables « dérivées » (ou « secondaires ») élaborées pour mieux correspondre aux exigences techniques du traitement statistique ainsi qu'aux exigences théoriques de la problématique. Ces variables sont dérivées au sens où elles résultent des variables primaires par recodage ou agrégation de plusieurs informations primaires. Ce sont les « vraies » variables sociologiques, celles directement liées à la problématique ou au questionnement théorique du sociologue.

2. VARIABLES QUALITATIVES ET VARIABLES QUANTITATIVES

Deux grands types de variables peuvent être distingués : les variables quantitatives, qui expriment des grandeurs quantifiables, et les variables qualitatives, qui reflètent des grandeurs non quantitatives, des « qualités ». En sociologie les secondes sont plus fréquentes que les premières : l'essentiel des informations est de nature qualitative. Ceci résulte de la nature des phénomènes analysés par le sociologue : les pratiques, les opinions, les représentations, les caractéristiques sociales, ou encore les attitudes s'expriment rarement à l'aide de variables quantitatives. Et il n'est pas rare que les quelques variables quantitatives soient recodées en variables qualitatives afin d'harmoniser le statut des variables et d'écarter l'illusion de précision que peuvent incarner les variables quantitatives.

La distinction entre variables quantitatives et qualitatives n'est pas anodine. Elle ne résulte pas d'un raffinement conceptuel inutile mais d'une contrainte technique forte : la nature des variables conditionne le type de méthodes d'analyse utilisables. Il est par exemple impossible de calculer un statut matrimonial moyen ou un diplôme moyen.

2.1 Variables quantitatives

Une variable quantitative permet d'exprimer une grandeur quantifiable c'est-à-dire une grandeur mesurable à l'aide d'une unité. C'est par exemple le cas de l'âge (exprimables en « années » ou en « mois »), du revenu (en euros ou en Yens) ou encore du nombre d'enfants. Une variable quantitative s'exprime à l'aide de nombres et ses diverses valeurs peuvent être numériquement comparées.

De manière générale, les sociologues utilisent des variables quantitatives dans deux grands types de situations. Premièrement, lorsqu'ils veulent exprimer des durées (âge, ancienneté d'une pratique, durée d'une expérience professionnelle, temps consacré à une activité, nombre d'années d'études, durée entre deux événements...), des valeurs monétaires (revenus, patrimoine, salaires, montant de l'argent de poche, dépenses, consommation, épargne...), des indicateurs de « volume » (nombre de livres lus, nombre d'enfants, taille du réseau amical...) ou des indicateurs d'« intensité » (fréquence d'une pratique culturelle...). Les variables synthétiques, que nous définirons plus loin et qui jouent un rôle central en sociologie quantitative,

relèvent également de cette catégorie : elles expriment grâce à un indicateur quantitatif la position d'un individu selon une grandeur sociologique – par exemple, son niveau de participation aux tâches ménagères, son niveau d'investissement sociale, son degré de « religiosité »...

Le second cas d'utilisation de variables quantitatives en sociologie est relatif aux situations où les sociologues ne travaillent pas sur des personnes, mais sur des entités collectives (par exemple des familles, ménages, associations, communes, entreprises...). Dans ce cas, ces collectifs peuvent être caractérisés par des variables quantitatives exprimant des parts ou des taux : part des individus de sexe masculin ; taux de redoublement ; part des plus de 65 ans ; part de ceux déclarant aimer la musique Rap ou RnB ; probabilité des enfants des différents groupes sociaux d'accéder à une grande école... Dans ce cas, on parle parfois de *données agrégées* car pour obtenir des caractéristiques relatives à des groupes, il est souvent nécessaire d'agréger les réponses individuelles.

2.2 Variables qualitatives

Les grandeurs non quantifiables sont celles qui ne peuvent pas s'exprimer en unités : ces modalités marquent des différences qui ne sont pas des différences numériques mais des différences de nature. Le diplôme, le sexe, la catégorie sociale, les sympathies politiques, le titre du dernier ouvrage lu, le statut matrimonial ou encore la couleur des yeux sont non quantifiables : elles s'expriment grâce à des variables qualitatives. Les modalités de ces variables ne sont pas comparables quantitativement : il n'existe aucune mesure commune de la modalité « marié » et de la modalité « divorcé » de la variable « statut matrimonial ».

Sont également considérées comme qualitatives les variables qui sont fondamentalement quantitatives mais que le sociologue utilise sous une forme recodée, avec des modalités qui correspondent à des classes. L'âge biologique est une variable quantitative mais elle est presque exclusivement utilisée sous la forme d'une variable qualitative définie à partir de classes d'âge : par exemple [18-25 ans] ; [26-30 ans] ; [31-40 ans] ; [41-55 ans] ; [56 ans et plus].

Parmi les variables qualitatives, il est possible de distinguer les variables à modalités ordonnables et celles à modalités non ordonnables. Comme leur nom l'indique, les modalités ordonnables peuvent être classées, hiérarchisées : c'est notamment le cas de toutes les variables dont les modalités sont semblables à « Tout à fait, assez, peu, pas du tout » ou « Très souvent, assez souvent, de temps en temps, rarement, jamais ». C'est aussi le

cas de toutes les variables fondamentalement quantitatives mais qui sont codées selon une échelle comme dans l'exemple suivant :

« Au cours de la dernière année, combien de livres avez-vous acheté ?

1. Aucun
2. Un ou deux livres
3. Entre 3 et 10 livres
4. Entre 11 et 30 livres
5. Plus de 30 livres »

Il est également possible de considérer que les variables « diplôme » voire « opinion politique » sont ordonnables : les diplômes peuvent être classés selon un principe de hiérarchie scolaire et de nombre d'années d'études ; les opinions politiques peuvent être classées en fonction de l'axe gauche-droite (à condition d'ignorer les difficultés concernant les apolitiques ou les écologistes). La catégorie sociale donne également lieu à un classement dans beaucoup de travaux sociologiques : catégories sociales supérieures, intermédiaires ou populaires...

Une variable qualitative peut être simple (lorsqu'elle reflète une seule information), multiple (lorsqu'elle reflète plusieurs informations en même temps) ou ordonnées (lorsqu'elle reflète plusieurs informations classées par ordre). La question « Quelles sont vos trois stations de radio préférées ? » constitue une variable multiple. S'il est, en plus, demandé de classer ces trois stations de radio préférées, elle devient une variable multiple ordonnée.

3. De la nécessité de recoder les variables

Le travail de recodage résulte de deux nécessités. L'une d'entre elles correspond à des contraintes statistiques et techniques : 1) certaines réponses, notamment les réponses aux questions ouvertes, doivent être recodées de manière à être exploitables dans une perspective quantitative ; 2) certaines modalités de réponses sont rarement choisies et doivent donc être regroupées car les effectifs ne permettent pas de les analyser en tant que telles ; 3) enfin, il est parfois nécessaire, pour pouvoir utiliser certaines méthodes statistiques, de diminuer le nombre de modalités des variables (c'est le cas dans les analyses factorielles).

La seconde nécessité correspond aux exigences et choix théoriques : elle résulte de la problématique sociologique choisie. Recoder une variable, c'est préparer les données de façon à les rendre adéquates à la problématique. Cette

dernière affirmation est essentielle : en dehors des contraintes techniques signalées ci-dessus, le recodage d'une variable doit être réalisé en fonction d'un questionnement et non de présupposés extérieurs à la problématique.

Il est donc faux de croire que le recodage est une simple opération technique. Il s'agit d'une opération théorique, visant à rendre les variables les plus adéquates possibles à la problématique et aux notions en œuvre dans celle-ci. Bien recoder les variables est un impératif pour conduire une bonne analyse sociologique.

3.1 Techniques de recodage 1 : regrouper des modalités

Considérons la question suivante, adressée à des titulaires du baccalauréat :

Quelles études avez-vous poursuivies après votre baccalauréat ?

a) Aucune, arrêt des études	f) Faculté de droit
b) Classes préparatoires	g) Autre filière universitaire
c) IUT	h) École d'infirmières
d) BTS	i) École d'architecture
e) Faculté de médecine ou de pharmacie	j)...

Il y a au moins trois manières de recoder cette variable, selon qu'on s'intéresse à l'opposition entre ceux qui ont poursuivi des études post-bac et ceux qui ont arrêté ; à l'opposition entre ceux qui ont engagé des études courtes (IUT, BTS...) et ceux ayant débuté des cursus longs (médecine, classes préparatoires) ; ou à l'opposition entre les filières sélectives (classes préparatoires, IUT, médecine, pharmacie...) et filières moins sélectives (filière universitaire hors médecine, pharmacie et droit...). C'est la problématique et la question théorique posée au traitement statistique (par exemple un tableau croisé) utilisant la variable qui vont déterminer la nature du recodage, en l'occurrence du regroupement de modalités.

3.2 Techniques de recodage 2 : simplifier les variables multiples

L'analyse des variables multiples et ordonnées est parfois plus facile si elles sont transformées en variables qualitatives simples. Il est par exemple possible de

transformer une variable ordonnée en une variable multiple en ne retenant que les modalités choisies par les enquêtés et en écartant l'ordre indiqué. Et il est possible de transformer une variable multiple en une série de variables dites indicatrices : à chaque modalité *M* de la variable multiple est associée une variable indicatrice dont les modalités sont « a choisi » et « n'a pas choisi » la modalité *M*.

Il est également parfois utile de transformer une variable multiple en une simple variable quantitative comptant le nombre de modalités choisies par chaque enquêté.

3.3 Techniques de recodage 3 : simplifier les variables quantitatives

Le recodage des variables quantitatives est souvent indispensable. Il y a au moins deux raisons à cela. Il est, d'une part, commode voire impératif de disposer de variables ayant toutes un statut identique : la plupart des variables manipulées par les sociologues étant des variables qualitatives, il est commode de recoder les quelques variables quantitatives en variables qualitatives. Cette remarque ne s'applique évidemment pas aux quelques situations où l'essentiel des variables sont quantitatives, notamment dans les travaux de socio-démographie, de socio-économie, ou lorsque le sociologue travaille sur des collectifs (voir le chap. 2, § 2).

Recoder une variable quantitative revient à définir les bornes (ou frontières) des diverses catégories (appelées « classes »).

Il existe trois principes généraux de recodage d'une variable quantitative. Le premier principe est un principe « esthétique » ou « mathématique » : les diverses valeurs de la variable sont regroupées en tranches d'égale amplitude et dont les bornes sont « naturelles ». Selon ce principe, la variable « âge » sera recodée en tranches de 5 ou 10 ans, avec des frontières « rondes » : [10-20 ans] ; [21-30 ans] ; [31-40 ans]... Ce principe semble être le plus naturel et est d'usage très fréquent (notamment en démographie et dans les enquêtes très générales) mais il n'est pas nécessairement le plus pertinent ni toujours le plus adéquat aux données dont dispose le sociologue. Les deux autres modes de recodage répondent davantage, de ce point de vue, aux exigences du travail sociologique.

Le deuxième principe de codage est de nature « statistique » et vise à assurer que les catégories créées regroupent un nombre suffisant d'individus. Le sociologue essaie de trouver un compromis entre des catégories (ou classes) regroupant un trop grand nombre d'individus (et donc trop grossières

et tentant à « écraser » les éventuelles différences entre individus) et des catégories regroupant un trop petit nombre d'individus (rendant ainsi impossible ou illusoire leur analyse statistique). Une solution « optimale » consiste à créer des classes équilibrées, c'est-à-dire regroupant un nombre d'individus proche d'une classe à l'autre. Certains logiciels permettent de déterminer automatiquement les classes statistiquement équilibrées. Sinon, il faut procéder par tâtonnement, en essayant plusieurs configurations.

Le troisième principe de recodage est de nature plus sociologique et vise à assurer que les catégories créées correspondent à des situations sociologiques homogènes, similaires. Ainsi, un sociologue travaillant sur les transformations induites par l'arrivée d'un premier enfant dans une famille devrait concevoir les différentes classes de la variable « âge » en fonction de son objet : si la taille de l'échantillon le permet, il devra concevoir des classes d'âge fines autour de l'âge moyen d'arrivée du premier enfant (entre 28 et 30 ans), quitte à concevoir des classes plus vastes pour les âges éloignés de cet âge moyen.

En pratique, c'est au sociologue de trouver un compromis raisonnable et acceptable du point de vue statistique et sociologique : le recodage d'une variable quantitative doit respecter le principe statistique, sans pour autant sacrifier l'exigence du sens sociologique de la variable. Le critère esthétique ou mathématique est plus superflu mais peut malgré tout entrer en ligne de compte pour rendre les résultats plus pédagogiques (puisque plus familiers et plus simples en apparence).

3.4 Techniques de recodage 4 : coder les matériaux qualitatifs

Devant un matériau de nature qualitative (des lettres, les images, des textes... voire des entretiens), le sociologue doit commencer par déterminer quelles sont les informations à retenir : quelles sont les données pertinentes pour sa problématique ? Une fois ces choix opérés, il doit coder les données selon une grille standardisée. Nous avons déjà présenté (chap. 1, § 3) le cas de codage de lettres ou de sources vidéos. Considérons ici un autre exemple : le codage de petites annonces matrimoniales parues dans le *Chasseur français*[1]. Ces petites annonces présentent des formes trop hétérogènes pour être analysables sans codage préalable. À côté du sexe et de l'âge de l'annonceur, il existe bien

1. François de Singly, « Les manœuvres de séduction : une analyse des petites annonces matrimoniales », *Revue française de sociologie*, 1984, XXV, 4, p. 523-559.

d'autres caractéristiques méritant d'être analysées et recodées : le nombre de mots de l'annonce, la présence d'enfants, le verbe formant jonction entre l'offre et la demande (« rencontrerait, épouserait, cherche... »), le nombre d'éléments corporels cités, la présence de référence économique, les références morales ou culturelles, la présence de qualificatifs d'excellence physique (« bien physiquement, joli, beau... ») ou encore la présence de qualificatifs d'excellence sociale (« belle situation, grande propriété... »). En tout, l'auteur a repéré et codé 78 traits dans son corpus d'annonces – certains qualitatifs, d'autres quantitatifs. Ce travail lui permet d'appréhender les processus par lesquels « un individu tente de faire reconnaître sa valeur sociale en mettant en scène ses richesses les plus propres à séduire ». Cet exemple illustre bien un principe : il faut faire feu de tout bois et longuement réfléchir aux informations méritant d'être recodées. Même les matériaux apparemment pauvres (ici des annonces de quelques lignes) peuvent faire l'objet de codages précis et nombreux (ici 78 critères distinctifs ont été repérés).

Les réponses aux questions ouvertes (par exemple « Quels sont les titres des films que vous avez vus au cinéma au cours du dernier mois ? ») constituent un cas fréquent de variables nécessitant ce type de travail de codage[1].

3.5 Techniques de recodage 5 : combiner les variables

Afin de simplifier le travail d'analyse et de croisement, il est souvent utile de concevoir des variables combinant deux variables primaires. Les modalités de la nouvelle variable sont obtenues par combinaison des modalités des deux variables primaires. Cette technique est particulièrement utile lorsque l'analyse conduit à tenir compte de deux variables contextuelles ou explicatives en même temps. Il est par exemple fréquent de recourir à une variable combinant à la fois une information sur le sexe et une information sur l'âge[2] :

Variable âge × sexe

1. Homme de 18 à 34 ans

1. Il est parfois possible de recourir à des outils d'analyse textuelle : voir Pascal Marchand, *L'Analyse du discours assistée par ordinateur*, Paris, Armand Colin, 1998 ; Ludovic Lebart, André Salem, *Statistique textuelle*, Paris, Dunod, 1994 (épuisé mais consultable sur le site <http://ses.telecom-paristech.fr/lebart/>).
2. On pourra prendre soin de réfléchir à l'ordre avec lequel on croise les variables : dans l'exemple, le sexe vient avant l'âge et la variable reflète des groupes de sexe découpés selon l'âge. L'inversion des rôles fournit une variable davantage structurée par l'âge.

2. Homme de 35 à 59 ans
3. Homme de plus de 60 ans
4. Femme de 18 à 34 ans
5. Femme de 35 à 59 ans
6. Femme de plus de 60 ans

Cette technique est également utile pour rassembler deux informations qui vont naturellement ensemble mais qui font l'objet de deux questions différentes dans le questionnaire. Les questions « Quelle est votre religion ? » et « Êtes-vous pratiquant(e) ? » peuvent être assemblées de la manière suivante :

1. Sans religion
2. Catholique non pratiquant
3. Catholique pratiquant
4. Protestant non pratiquant
5. Protestant pratiquant
6. Musulman non pratiquant
7. Musulman pratiquant
8. *etc.*

Le nombre de modalités de la nouvelle variable est égal au produit du nombre de modalités de chacune des questions : il peut donc être élevé et rendre nécessaire un nouveau recodage pour regrouper des modalités (notamment celles qui sont rares).

4. Passer des variables aux indicateurs théoriques : les variables synthétiques

Nous avons souligné la nécessité de recoder les informations recueillies pour les ajuster à la problématique et au questionnement théorique. Mais ce premier travail sur les variables ne suffit pas : il est souvent nécessaire de concevoir, à partir des réponses aux questions, de nouvelles variables incarnant les concepts et notions utilisés en les opérationnalisant. Ces variables sont des *variables* ou *indicateurs synthétiques* : elles rassemblent (« synthétisent ») les informations issues de diverses questions liées à un concept ou une notion.

Les notions ainsi opérationnalisées peuvent être abstraites et être issues de la théorie sociologique : c'est par exemple le cas des notions d'autonomie, d'individualisation, d'investissement scolaire, de proximité sociale ou encore d'intégration qui ne s'observent pas directement. Mais il peut également s'agir de notions moins abstraites dont l'objectivation passe nécessairement par plusieurs questions. Ainsi, plutôt que de demander « Lisez-vous beaucoup ? », il est préférable de poser plusieurs questions plus précises comme « Au cours du dernier mois, combien de romans avez-vous lus ? », « Combien de BD ? », « Combien d'essais ? », « Lisez-vous régulièrement un magazine ? », « Lisez-vous régulièrement un quotidien ? »... Ces diverses questions ne nous intéressent peut-être pas en tant que telles. Elles prennent sens dans la mesure où, prises ensemble, elles renseignent sur la pratique de lecture de l'enquêté. Mais, étant nombreuses, elles ne sont pas aisément utilisables dans les traitements statistiques. Il est dès lors utile de les rassembler pour constituer un indicateur synthétique d'intensité de la pratique de lecture.

Le nombre d'informations primaires intervenant dans la définition de la variable synthétique peut être très différent (de deux ou trois à quelques dizaines). Nous allons présenter les principales techniques permettant de construire et mettre au point de tels indicateurs synthétiques.

4.1 Créer des variables synthétiques par combinaison

La première technique, déjà entrevue précédemment, consiste à fusionner deux ou trois variables primaires en combinant leurs modalités. Imaginons travailler sur les pratiques de lecture et de « consommation » de livres et considérons par exemple les trois questions suivantes :

Q1. Au cours du dernier mois, avez-vous acheté des livres ?

1. Oui
2. Non

Q2. Au cours du dernier mois, avez-vous emprunté des livres en bibliothèque ?

1. Oui
2. Non

Q3. Au cours du dernier mois, avez-vous emprunté des livres à des proches, des amis, des connaissances... ?

1. Oui
2. Non

Il est possible de combiner ces trois questions pour construire la variable synthétique « Pratiques de l'achat et de l'emprunt de livres au cours du dernier mois » qui est un assez bon indicateur de la pratique livresque et qui peut être utile pour estimer la circulation et la manipulation des livres de manière indépendante de leur lecture :

1. N'a ni acheté ni emprunté
2. A acheté mais n'a pas emprunté
3. A acheté et a emprunté à des proches
4. A acheté et a emprunté en bibliothèque
5. A acheté et a emprunté à des proches et en bibliothèque
6. N'a pas acheté mais a emprunté en bibliothèque
7. N'a pas acheté mais a emprunté à des proches
8. N'a pas acheté mais a emprunté à des proches et en bibliothèque

Cette technique est notamment utilisée par Bernard Lahire dans son travail sur la *Culture des individus*[1] pour identifier le caractère consonant ou dissonant des pratiques et goûts culturels des individus. Ils sont dits dissonants s'ils mêlent à la fois des goûts et des pratiques très légitimes et peu légitimes ; ils sont consonants s'ils mêlent uniquement des aspects très légitimes (consonants légitimes) ou uniquement des aspects peu légitimes (consonants peu légitimes). Pour cela, la première étape est de classer, par simple recodage (regroupement de modalités), les différentes modalités de chaque variable décrivant les goûts et les pratiques (TV, musique, livres, visites, spectacles, cinéma...) selon leur degré de légitimité. Ainsi, par exemple, les préférences télévisuelles sont classées en trois modalités : peu légitimes (*Le Juste Prix*, *Tout est possible*, *Perdu de vue*...), très légitimes (*Bouillon de culture*, *Faut pas rêver*, *Les Mercredis de l'histoire*...) et mixtes. Par combinaison de ces variables, il est possible de qualifier chaque individu selon son profil : ses goûts et pratiques son dissonants en matière télévisuelle, livresque et cinématographique s'il combine, en ces matières, à la fois des modalités « peu légitimes » et « très légitimes ».

Cette technique de construction de variables synthétiques est seulement utilisable si le nombre de variables à combiner n'est pas trop élevé (deux ou trois, quatre au plus) et que le nombre de modalités de chacune de ces variables n'est également pas trop grand. Dans le cas contraire, la variable synthétique obtenue n'est pas commode d'utilisation puisque son nombre

1. Bernard Lahire, *La Culture des individus. Dissonances culturelles et distinction de soi*, Paris, La Découverte, 2004, chapitres 3 et 6.

de modalités est très élevé – et donc le nombre d'individus par modalité faible.

4.2 Créer des variables synthétiques par calcul de scores

Une seconde technique, qui est certainement la plus utilisée et la plus facile à mettre en œuvre, est de calculer des variables-scores. Le principe est le suivant : après avoir identifié la liste de toutes les variables utiles, on combine les variables ou certaines de leurs modalités. Cette combinaison de variables est différente selon que nous avons à faire à des variables qualitatives ou des variables quantitatives.

S'il s'agit de variables qualitatives, on affecte des notes (généralement des nombres entiers 0, 1 ou 2) à chacune des modalités des variables de cette liste puis, pour chaque individu, on compte son score ou sa note finale. Cette variable est par définition quantitative : elle doit être recodée car il est bien difficile de donner un sens à chacune des valeurs de cette variable. Il est usuel de créer un nombre réduit de classes (3, 4 ou 5 environ) : un échelonnement des comportements ou des situations individuelles en 3, 4 ou 5 niveaux suffit en général. Les modalités correspondent alors à des intensités : « très faible », « faible », « assez élevé », « très élevé »... Le principe n'est pas difficile à mettre en œuvre. Et il est très efficace pour créer des variables objectivant des notions plus ou moins abstraites.

Un cas simple et très fréquent de variable-score est celui où le sociologue souhaite calculer un nombre indiquant une intensité ou une diversité. Considérons par exemple la question « Parmi les activités physiques et sportives suivantes, indiquez celles que vous pratiquez au moins une fois par an ? Natation ; Jogging ; Vélo ; Randonnée ; Danse ; Gymnastique... ». Il peut être intéressant de calculer le nombre d'activités pratiquées, qui peut être interprété comme l'intensité de pratiques physiques et sportives d'un individu. Pour cela, il suffit d'attribuer la note 1 à toutes les activités puis de compter le score de chaque individu. Il peut également être intéressant de calculer la diversité des pratiques en distinguant les pratiques sportives individuelles, collectives et en comptant le nombre d'activités de nature différente pratiquées par chaque enquêté...

Une autre situation classique de construction d'une variable-score est le cas de l'opérationnalisation d'un concept ou d'une notion abstraite. Considérons par exemple notre recherche portant sur les usages du téléphone portable au

sein des couples, où nous nous intéressions notamment à la question de l'individualisation du portable[1]. Le portable peut en effet être utilisé par un individu à titre purement privé et individuel, sans que son conjoint n'intervienne. Cette notion s'oppose à celle de partage ou de collectivisation du portable.

Afin de préciser et d'opérationnaliser cette notion d'« individualisation du portable », nous avons utilisé les réponses à quatre questions du questionnaire : « Arrive-t-il que le conjoint réponde à votre place avec votre portable ? », « Au cours de la dernière semaine, votre conjoint a-t-il emprunté votre portable ? », « Au cours de la dernière semaine, avez-vous reçu des appels sur votre portable pour votre conjoint ? », « Votre conjoint connaît-il le code PIN de votre portable ? ». Notre indice général d'individualisation synthétise les réponses à ces quatre questions : pour chaque question où il a répondu négativement, un individu se voit attribuer un point. Ainsi, chaque individu est caractérisé par un « score » résumant le degré d'individualisme de son portable. La variable ainsi créée constitue une échelle d'individualisation du portable. De nature quantitative, elle varie de 0 à 4 : les scores 0 et 1 correspondant à un « très faible » ou « faible individualisme » du portable, le score 2 à un « individualisme moyen », le score 3 à un « individualisme assez fort » et le score 4 à un « très fort individualisme ».

Un autre exemple est celui mis en œuvre par Alain Girard dans son travail sur *Le Choix du conjoint* (1974). Pour étudier la « distance » ou inversement la « proximité » entre deux conjoints, c'est-à-dire leur dissemblance ou leur ressemblance sociale, culturelle et géographique, Alain Girard construit un indice global en retenant douze variables caractérisant les deux conjoints[2] :

- leur nationalité ;
- la taille de leur commune de naissance ;
- la situation géographique de leur commune de naissance ;
- leur niveau d'études ;
- leur religion ;
- la taille de leur commune de résidence au moment du mariage ;
- la situation géographique de leur commune de résidence lors du mariage ;

1. Olivier Martin, François de Singly, « Le téléphone portable dans la vie conjugale : retrouver un territoire pour soi ou maintenir le lien conjugal ? », *Réseaux*, vol. 20, 2002, n° 112-113, p. 211-248.
2. Alain Girard, *Le Choix du conjoint. Une enquête psycho-sociologique en France*, Paris, PUF-INED (nouvelle édition), 1974, p. 87-94. Le choix des variables participant à la définition de l'indicateur est critiquable, mais ce n'est pas l'essentiel ici.

– le nombre de localités habitées depuis leur naissance ;

– leurs professions ;

– la nationalité du père de chaque conjoint ;

– la profession du père de chaque conjoint ;

– la profession actuelle du mari et celle de son beau-père.

Selon la plus ou moins grande différence au sein de chacune des variables, Girard affecte une note variant de 1 à 7 : 1 si les deux conjoints sont très différents ; 7 s'ils sont semblables[1]. Ainsi, pour la variable « Niveau d'étude » divisée en sept modalités hiérarchisées en « degrés » (pas d'études, études primaires sans CEP, études primaires avec CEP, études techniques, études primaires, études secondaires, études supérieures), les conjoints sont affectés de la note 7 s'ils ont exactement le même niveau d'étude, de la note 6 si leurs niveaux d'étude diffèrent d'un degré..., de la note 1 si leurs niveaux d'étude diffèrent de six degrés.

Au final, les individus obtiennent une note (ou score) variant du minimum 12 au maximum 84 : d'une faible proximité à une proximité très forte. Il s'agit d'une variable quantitative que Girard regroupe en classes.

D'un point de vue technique, deux questions peuvent se poser lors de la création d'une variable-score. La première question est celle du choix des variables rentrant dans la composition de la variable synthétique : comment sélectionner les variables ? La principale réponse est d'ordre théorique : il faut inclure les variables issues des questions conçues et utilisées comme des indicateurs de la notion théorique au moment de la mise en point du questionnaire. Il est possible, mais pas indispensable, de compléter cette réponse par un second critère : il faut étudier les relations (par calcul des corrélations ou test de l'indépendance) entre les variables et inclure celles qui sont « positivement liées », c'est-à-dire qui « vont dans le même sens ».

La seconde question est celle du choix des coefficients de pondération affectés à chacune des variables ou des modalités. Il n'existe aucun critère indiscutable pour justifier le choix de la valeur de coefficients de pondération. Les sociologues recourent simplement à leur bon sens, c'est-à-dire choisissent les valeurs en fonction de la pertinence de la variable ou de la modalité et de sa capacité à exprimer une notion théorique. L'expérience

1. Girard interprète son indicateur comme une distance. Mais il s'agit plutôt d'un indicateur de proximité : une valeur élevée de l'indicateur étant synonyme d'une forte proximité entre les conjoints ; une valeur faible étant synonyme de fortes différences sociales entre les individus.

tend de toute façon à montrer que le choix des coefficients n'est pas crucial. Le plus souvent, on se contente d'affecter des notes simples : 0 et 1 ; voire 0, 1 et 2...

La qualité de la variable synthétique dépend bien davantage de la pertinence des variables qui rentrent dans sa composition. Cette qualité dépend aussi du nombre de ces « variables pertinentes » : *a priori*, plus leur nombre est élevé, plus la variable synthétique aura un sens incontestable et robuste.

4.3 Créer des variables synthétiques à partir de variables quantitatives

Dans le cas du calcul de variables-scores, des points sont attribués aux modalités puis additionnés car, les variables étant qualitatives, il n'est pas possible de combiner numériquement leurs modalités. Dans le cas où les variables sont quantitatives, il est possible de combiner directement leurs valeurs par addition, soustraction, division... ou toute autre opération mathématique[1]. Le seul critère pour juger du bien-fondé de l'indicateur calculé est la signification ou le sens qu'il est possible de lui attribuer.

Par exemple, pour étudier le « lien de germanité à l'âge adulte » et notamment l'homophilie de sexe (à l'âge adulte, a-t-on davantage tendance à fréquenter les individus de même sexe que soi parmi les membres de sa fratrie ?), des sociologues[2] ont eu recours à des indicateurs d'homophilie définis comme la différence entre le nombre de rencontres avec des germains de même sexe au cours des 12 derniers mois et le nombre de rencontres avec des germains de sexe différent au cours des 12 derniers mois. Cet indicateur, de définition et d'interprétation simples, permet d'objectiver cette notion d'homophilie de sexe : les valeurs élevées de l'indicateur sont des signes d'homophilie prononcée ; les valeurs faibles (négatives) sont des signes d'absence d'homophilie, voire d'hétérophilie de sexe.

De manière comparable, étudiant les activités et les loisirs et notamment les disparités entre les hommes et les femmes, Alain Chenu et Nicolas Herpin

1. Quitte à pondérer ou normaliser les variables initiales si leurs valeurs et leurs variabilités sont trop hétérogènes (en d'autres termes, si elles prennent des gammes de valeurs très différentes : l'une variant de 0 à 1 ; l'autre variant de 1 000 à 10 000 par exemple).
2. Emmanuelle Crener, Jean-Hugues Déchaux, Nicolas Herpin, « Le lien de germanité à l'âge adulte. Une approche par l'étude des fréquentations », *Revue française de sociologie*, vol. 41, n° 2, 2000, pp. 211-239.

ont défini un indicateur du caractère plutôt masculin ou plutôt féminin d'une activité[1]. Pour cela, ils ont considéré, pour chaque activité, le temps moyen D_h passé par les hommes dans cette activité et le temps moyen D_f passé par les femmes. Leur indicateur est alors défini ainsi :

$$I = 200 \times \frac{D_f}{D_f + D_h} - 100$$

Cet indicateur peut varier entre -100 et 100 : il prend la valeur 100 pour une activité exclusivement féminine, la valeur -100 pour une activité exclusivement masculine, la valeur 0 pour une activité indifféremment masculine et féminine. Cet indicateur leur permet d'identifier facilement les dominantes plutôt féminines ou masculines de diverses activités : la lecture, la promenade, le sport, le bricolage, les courses, le soin des enfants... Par exemple, l'indicateur vaut environ 50 en ce qui concerne la cuisine, le linge et le ménage : ces activités sont nettement féminines (les femmes y passent trois fois plus de temps que les hommes) ; et l'indicateur vaut presque 100 en ce qui concerne la couture (cette activité est quasi exclusivement féminine). Le calcul de ces indicateurs à deux dates différentes permet par ailleurs de déterminer si une activité se féminise ou non au cours du temps, ou si elle perd progressivement de sa dominante féminine...

4.4 Créer des variables synthétiques par analyse factorielle

Afin de diminuer l'arbitraire dans le choix des variables et de leur poids (coefficient de pondération) dans la définition d'un indicateur synthétique, il est possible de recourir à des méthodes statistiques dites multidimensionnelles c'est-à-dire destinées à analyser un grand nombre de variables en même temps. Nous les présenterons en détail au chapitre 4. Retenons pour l'instant que certaines d'entre elles – notamment les méthodes factorielles – permettent de construire de nouvelles variables qui soient des combinaisons de variables et qui restituent au mieux les différents liens entre ces variables. Ces nouvelles variables sont appelées des « axes », des « facteurs » ou des « dimensions ».

1. Alain Chenu et Nicolas Herpin, « Une pause dans la marche vers la civilisation des loisirs », *Économie et statistique*, n° 352-353, 2002, p. 15-37.

Dans sa recherche sur la carrière scolaire des enfants issus de l'immigration[1], Philippe Cibois formule l'hypothèse que si ces enfants réussissent mieux (à caractéristiques sociales identiques) que les enfants de parents français, c'est en raison de la force de leur projet migratoire et de leurs attentes vis-à-vis du système scolaire. Les familles immigrées ont, selon l'auteur, des attentes qui se caractérisent par une bonne volonté scolaire, c'est-à-dire « par un ensemble de comportements de respect des consignes données par l'école dans le comportement scolaire et hors école des enfants ». Pour construire son indicateur de « bonne volonté scolaire », Philippe Cibois utilise des modalités à 15 questions en ne retenant que les modalités considérées comme des indices de bonne volonté scolaire. Par exemple : « avoir préparé son cartable la veille avant de se coucher », « avoir préparé son cartable la veille avant le repas », « n'oublie jamais ou rarement un livre ou un cahier à la maison », « estime qu'arriver en retard à l'école est grave »... Mais, plutôt que de construire un indicateur « à la main » en choisissant une pondération de manière arbitraire, l'auteur réalise une analyse des correspondances qui lui permet d'obtenir une nouvelle variable (un facteur) opposant, de manière synthétique et cohérente, la bonne volonté scolaire à une attitude qu'il qualifie de « décontractée » vis-à-vis de l'école[2].

1. Philippe Cibois, « La bonne volonté scolaire. Expliquer la carrière scolaire d'élèves issus de l'immigration », *in* Philippe Blanchard et Thomas Ribémont (dir.), *Méthodes et outils des sciences sociales. Innovation et renouvellement*, L'Harmattan, 2002, p. 111-126. Pour un autre exemple, voir Olivier Galland, Yannick Lemel et Jean-François Tchernia, « Les valeurs en France », *Données sociales*, Paris, INSEE, 2002, p. 559-564.
2. En fait, cette analyse a une fonction exploratoire (voir cette notion dans notre conclusion) : elle permet à l'auteur de construire son indicateur en toute connaissance de cause.