SCIENCES HUMAINES

Introduction à la lecture des tableaux statistiques

par Jean-Paul Grémy
Professeur honoraire à l'université de Paris V

Ce manuel propose quelques principes empiriques destinés à faciliter la lecture des tableaux statistiques que l'on rencontre dans les sciences sociales. À cette fin, il présente plusieurs types différents de tableaux tirés de la littérature sociologique, et montre comment il est possible d'en extraire les informations essentielles sans utiliser d'autres outils mathématiques que les quatre opérations arithmétiques. Les exemples sont présentés par ordre de complexité croissante. Ils sont l'occasion de rappeler les notions de base indispensables pour lire et analyser les données quantitatives (niveaux de mesure, liaison entre variables, interaction, etc.), mais en les considérant seulement du point de vue des services qu'ils peuvent rendre au sociologue ; tout formalisme mathématique en a donc été exclu. C'est pour introduire ces notions d'une manière progressive que ce manuel accorde une place qui pourrait sembler trop importante aux tableaux les plus simples que l'on puisse trouver : les distributions à une seule dimension.

Les règles pour lire un tableau statistique qui y sont énoncées sont tirées principalement de l'expérience personnelle de l'auteur, tant dans la recherche (fondamentale et appliquée) que dans l'enseignement des méthodes des sciences sociales [1]. Aussi propose-t-il plutôt des conseils pratiques (voire des "recettes") que des connaissances théoriques (que l'on peut trouver d'ailleurs dans tous les manuels de statistique).

À l'occasion de chacun de ces exercices commentés, le lecteur est invité à refaire lui-même sur les données les opérations (simples) qui lui sont présentées ; il s'aidera pour cela d'une calculette, et il utilisera si possible les ressources graphiques que les tableurs mettent à sa disposition. Comme, dans la plupart des cas, il reste encore de nombreuses informations à extraire du tableau présenté, il pourra s'il le désire pousser plus loin l'analyse esquissée. Ensuite, principalement dans les derniers exemples (tableaux à plus de deux dimensions), il pourra s'inspirer des démarches qu'il aura assimilées pour définir sa problématique personnelle et aborder à sa manière ces données complexes.

À l'issue de cet apprentissage, le lecteur aura découvert que, comme son nom l'indique, la *lecture* des tableaux statistiques est avant tout un exercice littéraire : il s'agit principalement de décrypter les intentions des auteurs des tableaux, et d'analyser la signification des variables et des catégories définies sur celles-ci. Cet exercice de lecture se complète d'un exercice de *rédaction*, puisqu'il faut ensuite exprimer en langage naturel ce que les chiffres nous ont appris. Plus encore que des connaissances statistiques (certes indispensables), ces deux types d'exercices requièrent la recherche du mot juste et de la précision dans l'expression.

[1] Ce manuel s'inspire, dans ses grandes lignes, d'un document pédagogique portant le même titre, qui a été expérimenté et diffusé antérieurement par le LEMTAS, laboratoire de l'Université René Descartes (Paris 5).

Chapitre 1. Distributions à une variable.

Les distributions à une seule variable constituent les exemples les plus simples de tableaux statistiques. Dans les opérations de dépouillement d'enquêtes, on les nomme également *marginaux* ou *tris à plat* (par opposition aux *tris croisés*, qui sont des distributions à deux variables ou plus). Malgré leur simplicité, ces tableaux sont d'une grande importance pour la lecture des données quantitatives. En effet, nous verrons plus loin que tout tableau à plusieurs dimensions peut en fait se réduire à une série de tableaux à une dimension ; savoir lire correctement un tableau à une dimension est par conséquent une étape indispensable pour lire des données plus complexes.

Nous donnons ci-après trois exemples de distributions à une variable. Nous verrons que chacune de ces distributions présente des propriétés différentes, autorisant des traitements statistiques différents. Nous profiterons de ces exemples pour rappeler ce que l'on entend par *distribution*, *variable*, *niveau de mesure*, *population*, et *échantillon*.

1.1. Premier exemple : la structure de la population active française en 1982.

Le tableau 1 est extrait d'un tableau plus général établi par l'INSEE à partir des données du trente et unième recensement de la population française, en 1982 [1]. Pour chacune des personnes recensées, on a recueilli diverses informations concernant ses activités professionnelles ; en particulier : est-elle active au moment de l'enquête ? Si oui : quelle est sa profession ? Les déclarations du répondant ont été relevées en toutes lettres par l'enquêteur. Elles ont ensuite été reprises et codées selon une nomenclature des professions établie par l'Institut National de la Statistique et des Études Économiques (INSEE). L'année 1982 marque un changement important pour tous les sociologues qui étudient les structures sociales en France : à l'occasion de ce recensement, l'INSEE inaugure une nouvelle classification des catégories socioprofessionnelles, rendue nécessaire à la fois par les modifications des conditions d'exercice et de la structure des professions, et par l'évolution des conceptions sur les classifications sociales. Cette nouvelle nomenclature (PCS) a en outre été conçue pour permettre les comparaisons diachroniques (variations dans le temps), en restant compatible avec la nomenclature précédente (CSP) [2].

a) Lecture du tableau.

On remarquera que ce tableau ne porte que sur la population *active*, qui ne représente en 1982 que 43 % de la population totale (alors de 54,273 millions d'individus). Comme la dernière ligne du tableau (avant la totalisation) est intitulée "chômeurs n'ayant jamais travaillé", on peut en induire que la population active inclut les demandeurs d'emploi, qu'ils aient déjà travaillé ou non ; les personnes actuellement au chômage sont alors classées selon la dernière profession exercée. Par contre, n'y figurent ni les élèves ou étudiants, ni les

[1] Extrait du tableau III, page 27, de : Desrosières, Alain, Thévenot, Laurent, *Les catégories socio-professionnelles*, Paris, Éditions La Découverte, 1988.
[2] Le livre d'où est tiré ce tableau est entièrement consacré à la présentation de cette nouvelle nomenclature, et à l'analyse sociologique des grandes catégories qu'elle distingue.

militaires du contingent, ni les inactifs ; cette dernière catégorie inclut les "femmes au foyer". Au cours de l'analyse du tableau, il faudra garder présente à l'esprit cette définition de la population active (dans certaines recherches, comme par exemple celles sur la mobilité sociale, on pourrait inclure les retraités dans les catégories professionnelles correspondant à leur ancienne activité).

Professions et catégories sociales (PCS)	Nombre (en milliers)	Proportion (pour mille actifs)
Agriculteurs exploitants	1 475	63 ‰
Artisans, commerçants et chefs d'entreprise	1 835	78 ‰
Cadres et professions intellectuelles supérieures	1 895	81 ‰
Professions intermédiaires	3 971	169 ‰
Employés	6 247	265 ‰
Ouvriers	7 749	329 ‰
Chômeurs n'ayant jamais travaillé	353	15 ‰
Population active	23 525	1 000 ‰

Tableau 1. Structure de la population active en 1982.

Ce tableau présente à la fois les effectifs bruts de chaque catégorie (arrondis au millier le plus proche), et la part relative de ces catégories dans l'ensemble de la population active (en "pour mille" ; on vérifiera que : 1 475 / 23 525 × 1 000 = 62,699 ≅ 63 ‰). Les proportions et les effectifs ont été arrondis de manière à faciliter la lecture ; comme il est de règle, le gain en clarté a été compensé par une (légère) perte en précision.

La PCS est généralement considérée comme une variable *nominale*, c'est à dire une variable qui ne permet guère que de *nommer* les individus qu'elle sert à décrire. Pour ce faire, elle affecte à chaque individu une étiquette, et une seule ; elle réalise par conséquent une *partition* de l'ensemble la population active. Il s'agit ici bien entendu d'une simplification de la réalité, puisque non seulement une même personne peut avoir exercé des professions très différentes au cours de sa carrière (mobilité professionnelle), mais il arrive même que certaines personnes occupent simultanément deux emplois distincts (comme par exemple ces agriculteurs qui sont en même temps ouvriers de l'industrie). Dans ce tableau, il s'agit donc de la profession *principale* exercée en 1982.

Outre la PCS, les variables nominales les plus couramment utilisées par les sociologues sont le sexe, la région d'habitation, la situation de famille. Dans les variables de ce type, l'ordre de présentation des différentes catégories est en principe indifférent : on peut modifier cet ordre sans modifier la signification de la variable (nous verrons sur les exemples suivants que ce n'est pas le cas pour toutes les variables). Sur le schéma 1, qui est la traduction graphique de ce tableau, les PCS sont classées par ordre d'effectifs croissant. Traduire un tableau de chiffres en un graphique (histogramme ou courbes) en rend souvent la lecture plus facile.

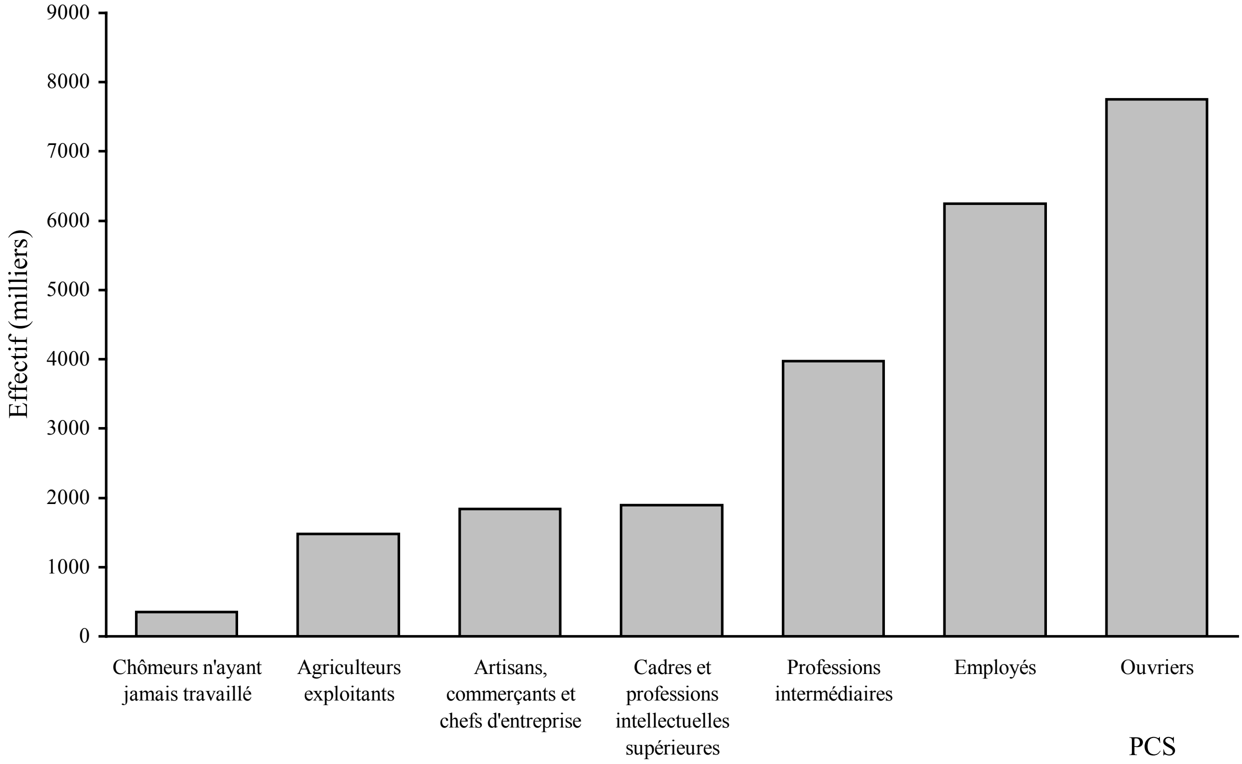


Schéma 1. Structure de la population active de la France en 1882.

Les seules informations que l'on peut tirer d'un tableau comme le tableau 1 proviennent de la comparaison des effectifs (ou des proportions) des diverses catégories. On voit immédiatement que la PCS qui est numériquement la plus importante est celle des ouvriers, qui représente près du tiers des actifs ; nous dirons que la catégorie "ouvriers" est le *mode* de la distribution. Les autres comparaisons possibles dépendent des intentions du chercheur : il peut souligner la faible part des agriculteurs (dont on sait par ailleurs que les effectifs diminuent régulièrement), ou l'importance relative des employés (dont on sait que les effectifs augmentent, sans pour autant avoir atteint le niveau des ouvriers) ; il peut également comparer chaque PCS à l'effectif moyen de celles-ci (3 862 milliers, en ne comptant pas les chômeurs n'ayant jamais travaillé). Enfin, il peut également, connaissant l'effectif de l'ensemble de la population recensée, calculer les proportions des actifs par PCS sur l'ensemble de la population. Le tableau 2, qui présente ces proportions, est exprimé en pourcentages, d'un usage plus courant que les "pour mille" ; l'effectif sur lequel ces pourcentages sont calculés est appelé la *base* des pourcentages.

Professions et catégories sociales (PCS)	Pourcentages de la population active	Pourcentages de l'ensemble de la population
Agriculteurs exploitants	6,3 %	2,7 %
Artisans, commerçants et chefs d'entreprise	7,8 %	3,4 %
Cadres et professions intellectuelles supérieures	8,1 %	3,5 %
Professions intermédiaires	16,9 %	7,3 %
Employés	26,5 %	11,5 %
Ouvriers	32,9 %	14,3 %
Chômeurs n'ayant jamais travaillé	1,5 %	0,7 %
Population active : 100,0 % = (en milliers)	23 525	43,3 %
Population totale : 100,0 % = (en milliers)	-	54 273

Tableau 2. Structure de la population active en 1982, par rapport à la population totale.

b) Comment interpréter ces données ?

Modifier l'ordre des catégories, rechercher le mode de la distribution, calculer des fréquences relatives ou des pourcentages sont des opérations usuelles pour décrire ou comparer des distributions sur des variables nominales. Ces opérations ne tiennent pas compte des *significations* attachées à chacune des catégories. Lorsque, pour analyser un tableau, on désire faire subir des transformations à une variable nominale, il n'est pas possible de faire abstraction de ces significations. Formellement, les propriétés de ce type de variable autorisent tous les regroupements de catégories possibles. Rien n'interdit par exemple de regrouper, dans une catégorie plus vaste, les chômeurs n'ayant jamais travaillé et les agriculteurs ; pourtant, du point de vue du sociologue, il n'est pas certain qu'un tel regroupement ait un sens. En effet, selon la perspective adoptée par le sociologue, une PCS peut ressembler plus à certaines PCS qu'à d'autres ; aux propriétés formelles de la *partition* que la variable PCS opère sur la population active s'ajoutent d'autres propriétés, qui dépendent du sens que le chercheur attribue à ces catégories.

Toute formalisation a un caractère fortement réducteur ; c'est le prix à payer pour effectuer des raisonnements et élaborer des constructions théoriques. La définition d'une échelle nominale n'échappe pas à cette règle. Mais, en dernier ressort, c'est au sociologue qu'il appartient de décider si cette formalisation lui convient, et d'en changer s'il l'estime nécessaire. Si celui-ci considère qu'il existe des ressemblances ou des oppositions entre les PCS, il devient légitime de simplifier le tableau en opérant des regroupement de catégories ayant un sens. On peut commencer par choisir de ne prendre en considération que les personnes ayant déjà travaillé, et donc susceptibles d'être caractérisées par une PCS ; cela réduit la population étudiée à 23 172 milliers d'actifs.

Ensuite, si l'on tient compte des significations attachées à chaque PCS, il est clair que chaque catégorie n'est pas différente de chacune des autres au même degré : par exemple, les employés sont plus proches des ouvriers, qui sont comme eux des salariés, que des agriculteurs exploitants. Il est par conséquent possible de regrouper certaines PCS sous un concept plus général (donc plus abstrait), exprimant une caractéristique qu'elles possèdent en commun. Une première distinction peut être faite selon qu'il s'agit de professions salariées ou d'indépendants : on regroupera ainsi les agriculteurs exploitants avec les artisans, commerçants et chefs d'entreprise, ce qui nous permettra d'indiquer que la population active compte 3 310 milliers d'indépendants (soit 14,3 %) contre 19 862 salariés (85,7 %). Toutefois, une telle affirmation devra être avancée avec prudence, puisque les professions libérales ont, dans le tableau 1, été confondues avec les cadres supérieurs [1]. Par ailleurs, ces regroupements permettent de calculer des pourcentages sur des totalisations partielles : par exemple, on pourra dire que les agriculteurs représentent 44,6 % des travailleurs indépendants, ou que 31,5 % des salariés sont des employés [2].

D'autres regroupements des PCS mériteraient d'être réalisés : par exemple selon le niveau de qualification, le degré de féminisation, le prestige social, ou encore le fait d'employer ou non de la main d'œuvre salariée. Les données que nous présentons ici sont trop sommaires pour permettre ces opérations ; dans l'ouvrage dont ce tableau a été extrait, on trouvera des classifications plus précises, permettant des regroupements de PCS plus intéressants du point de vue de la réflexion sociologique [3].

1.2. Deuxième exemple : l'opinion des Français sur le progrès scientifique.

Ces données sont extraites d'une enquête par sondage réalisée par l'Institut BVA pour l'association *Agorамétrie*. Cette association réalise régulièrement des enquêtes d'opinion sur des thèmes d'actualité, constituant un échantillon représentatif des grands thèmes abordés par les principaux médias. Ces enquêtes proposent aux personnes interrogées une série de phrases exprimant une opinion sur un sujet controversé, et leur demandent d'exprimer leur degré d'accord ou de désaccord avec chacune des phrases proposées, selon une gradation en cinq points. Le tableau 3 présente les réponses à l'une de ces questions (dont l'énoncé figure dans le tableau lui-même, ce qui favorise une meilleure interprétation des réponses)[4].

On peut ne pas approuver la formulation utilisée ici pour les degrés d'accord. Il est en effet plus habituel d'utiliser une gradation de la forme : "pas du tout / plutôt pas / plutôt / tout à fait" ; cette dernière gradation présente l'avantage d'obliger le répondant à prendre nettement

[1] Elles représentent 239 000 personnes (*op. cit.*, page 27), soit 1,0 % des actifs ayant exercé une profession ; en conséquence, l'affirmation précédente devrait être modifiée ainsi :"la population active compte 3 549 milliers d'indépendants (15,3 %) contre 19 623 salariés (84,7 %)".
[2] Ou, en tenant compte de la note précédente, que les agriculteurs représentent 41,6 % des indépendants, et les employés 31,8 % des salariés.
[3] Classification de la population active ayant un emploi en 31 catégories (page 27), et classification plus détaillée en 455 postes (pages 112-121).
[4] Tableau 16, page 34 de *Les structures de l'opinion fin1997. Analyse du champ des controverses médiatiques et de sa dynamique*, Paris, Agorамétrie, octobre 1998.

parti pour ou contre l'opinion présentée ; en contrepartie, elle présente le risque d'augmenter le nombre de "non réponses" (ou : "sans opinion"). La formulation adoptée par *Agoramétrie* semble avoir été assez bien comprise par les personnes interrogées, d'autant que la liste des réponses possibles leur est présentée par l'enquêteur sur un carton qui rend perceptible leur gradation (variable *ordinale*).

"Au nom du progrès scientifique on fait plus de mal que de bien"

	Effectif	Pourcentage
- "pas du tout d'accord"	200	8,9 %
- "pas tellement d'accord"	529	23,5 %
- "peut-être d'accord"	712	31,6 %
- "bien d'accord"	417	18,5 %
- "entièrement d'accord"	300	13,3 %
- non réponse	91	4,1 %

Tableau 3. Opinion des Français sur le progrès scientifique.

La population interrogée en 1998 est un *échantillon* de 2 249 personnes, représentatif de l'ensemble des Français de 18 ans et plus. Cela signifie que, pour des raisons de délais et de coûts, on a interrogé seulement une petite partie de la population française visée, mais en sélectionnant les personnes interrogées de telle sorte qu'elles constituent un "modèle réduit" de la population dans son ensemble. Si l'on vérifie la *représentativité* de l'échantillon, on doit en principe y trouver la même proportion de femmes, de moins de 20 ans, d'ouvriers, d'habitants de communes rurales, etc., que dans la population visée par l'enquête. Les enquêtes sur échantillon (*sondages*) n'apportent pas la même précision qu'un recensement. Usuellement, on calcule les *limites de confiance* des pourcentages présentés (ou, plus simplement, on les lit dans une table) ; ce sont les "fourchettes" des estimations électorales. Plus la taille de l'échantillon interrogé est grande, plus ces "fourchettes" sont resserrées (et donc plus les estimations sont précises).

"Au nom du progrès scientifique on fait plus de mal que de bien"	Estimation au seuil de P = 0,05		
	Limites inférieures	Pourcentages observés	Limites supérieures
- "pas du tout d'accord"	7,7 %	8,9 %	10,1 %
- "pas tellement d'accord"	21,7 %	23,5 %	25,3 %
- "peut-être d'accord"	29,7 %	31,6 %	33,5 %
- "bien d'accord"	16,9 %	18,5 %	20,1 %
- "entièrement d'accord"	11,9 %	13,3 %	14,7 %
- non réponse	3,3 %	4,1 %	4,9 %

Tableau 4. Limites de confiance des pourcentages du tableau 3.

Le tableau 4 donne un aperçu des taux de réponses que l'on aurait probablement obtenus en interrogeant la totalité des Français de 18 ans et plus. Ainsi, la proportion de ceux qui ne sont pas du tout d'accord avec l'opinion citée se situe vraisemblablement entre 7,7 % et 10,1 % des répondants ; la proportion de personnes entièrement d'accord, entre 11,9 % et 14,7 %.

Dans une variable nominale, tous les regroupements de catégories sont *a priori* possibles. Dans une variable *ordinale* comme celle-ci, on ne peut en principe regrouper que des catégories contiguës, de manière à conserver la propriété d'ordre strict sur les catégories. Les regroupements les plus usuels sont les pourcentages cumulés. On dira par exemple que 24,5 % des répondants ne sont pas d'accord avec l'opinion exprimée, ou encore que 33,3 % sont nettement d'accord, mais que cette proportion passe à 66,6 % si l'on compte les accords moins marqués (tableau 5).

D'autres regroupements entre catégories contiguës sont naturellement possibles : on peut par exemple juger utile de signaler que 76,8 % des répondants donnent une réponse modérée (autour de la médiane). Par contre, les regroupements de catégories non contiguës doivent être justifiés par la signification des réponses, et l'interprétation qu'on leur donne ; ainsi, on peut affirmer que les réponses extrêmes, dénotant une opinion bien arrêtée, ne sont données que par 23,2 % des personnes interrogées.

Que faire des "non réponses" (catégorie nominale) dans une échelle ordinale ? Faut-il faire l'hypothèse que ceux qui n'ont pas répondu étaient indifférents à l'opinion présentée, et auraient tout aussi bien choisir la réponse centrale ("peut-être d'accord") ? Ou encore supposer que les non répondants, s'ils avaient répondu, auraient fourni proportionnellement les mêmes réponses que les autres personnes interrogées ? Dans le premier cas, on aurait ajouté les 4,1 % de "non réponses" à la catégorie centrale ; dans le second cas, on les aurait réparti dans les cinq catégories de réponse proportionnellement à l'effectif de ces catégories.

En fait aucune de ces solutions n'est satisfaisante : les "non réponses" sont en elles-mêmes une réponse à part entière (nous reviendrons sur ce point au § 1.4 c). Toutefois, afin de pouvoir calculer les pourcentages cumulés du tableau 5, ce qui n'est possible qu'avec des catégories ordonnées, nous avons choisi d'éliminer les "non réponses" du tableau, et de ne prendre en considération que les 2 158 personnes ayant exprimé explicitement une opinion. Nous verrons (§ 1.4 c) que cela revient à adopter la seconde hypothèse énoncée ci-dessus.

"Au nom du progrès scientifique on fait plus de mal que de bien"	Pourcentages bruts	Pourcentages cumulés	
		en ordre croissant	en ordre décroissant
- "pas du tout d'accord"	9,3 %	9,3 %	100,0 %
- "pas tellement d'accord"	24,5 %	33,8 %	90,7 %
- "peut-être d'accord"	33,0 %	66,8 %	66,2 %
- "bien d'accord"	19,3 %	86,1 %	33,2 %
- "entièrement d'accord"	13,9 %	100,0 %	13,9 %

Tableau 5. Pourcentages bruts et cumulés pour les personnes ayant exprimé une opinion.

À partir des pourcentages bruts, on voit immédiatement que le *mode* de la distribution est la réponse "peut-être d'accord". Dans le cas d'une variable ordinale, on peut aussi rechercher la *médiane* de la distribution, c'est à dire la valeur de la variable qui correspond à l'individu de rang médian (celui qui a autant d'individus avant lui qu'après lui). Si les 2 158 personnes qui ont répondu à cette question pouvaient être classées les unes par rapport aux autres en fonction de leur degré d'approbation avec cette opinion, le rang médian correspondrait à la 1 079 ième ou la 1080 ième personne ; mais comme ici les réponses possibles ne comportent que cinq degrés, le repérage de la médiane ne peut être qu'approximatif. Les pourcentages cumulés nous apprennent que la médiane est également la réponse "peut-être d'accord". Cela tient à la forme symétrique de la distribution.

Lorsque l'on regroupe des catégories, ou lorsque l'on calcule des pourcentages cumulés (ce qui est une forme particulière de regroupement), on n'obtient pas les limites de confiance de ces regroupements en additionnant simplement les limites de chacune des catégories regroupées. Il faut recalculer (ou lire dans les tables) les limites correspondant à ces pourcentages cumulés (tableau 6).

"Au nom du progrès scientifique on fait plus de mal que de bien"	Estimation au seuil de P = 0,05		
	Limites inférieures	Pourcentages cumulés	Limites supérieures
- "pas du tout d'accord"	8,1 %	9,3 %	10,5 %
- "pas tellement d'accord"	31,8 %	33,8 %	35,8 %
- "peut-être d'accord"	64,8 %	66,8 %	68,8 %
- "bien d'accord"	84,6 %	86,1 %	87,6 %
- "entièrement d'accord"	-	100,0 %	-

Tableau 6. Limites de confiance des pourcentages croissants du tableau 5.

1.3. Troisième exemple : la répartition par âge des élèves de troisième générale en 1997-1998.

Le tableau 7 présente les effectifs par âge des 685 208 élèves de troisième générale de l'enseignement public et de l'enseignement privé, pour l'année scolaire 1997-1998 [1]. Par conséquent, ne figurent dans ce tableau ni les élèves de troisième d'insertion (qui accueillent les élèves en difficulté et sont ouvertes sur le monde de l'entreprise), ni ceux de troisième technologique (qui associent aux enseignements généraux une formation technologique importante). Comme l'âge présente toutes les caractéristiques d'une variable ordinale (ordre strict sur les catégories), nous avons ajouté au tableau original trois colonnes, présentant les pourcentages (bruts et cumulés) correspondants aux effectifs ; on aurait pu également y faire figurer les effectifs cumulés, en ordre croissant (nombre d'élèves ayant au plus x ans) et décroissant (nombre d'élèves ayant au moins y ans).

Âge	Effectifs	% bruts	% croissants	% décroissants
- 13 ans et moins	22 953	3,4 %	3,4 %	100,0 %
- 14 ans	**390 207**	**56,9 %**	**60,3 %**	**96,9 %**
- 15 ans	210 791	30,8 %	91,1 %	39,7 %
- 16 ans et plus	61 257	8,9 %	100,0 %	8,9 %
Total	685 208	100,0 %	-	-

Tableau 7. Élèves de troisième générale en 1997-1998, selon l'âge.

[1] Extrait du tableau 1, page 87, de *Repères & références statistiques sur les enseignements et la formation*, Ministère de l'Éducation nationale, de la Recherche et de la Technologie, Paris, 1998.

On lit immédiatement sur ce tableau que le *mode* de la distribution est 14 ans ; quant à la *médiane*, elle se situe également à l'intérieur de cette catégorie. Dans le tableau original, la ligne "14 ans" est en caractères gras ; elle correspond à ce que l'on appelle l'"âge théorique normal", qui caractérise les "élèves à l'heure". "Un élève est dit 'à l'heure' lorsqu'il est entré en cours préparatoire (CP) à 6 ans et qu'il parcourt sa scolarité sans redoublement ni saut de classe" [1]. Nous lisons sur ce tableau que 56,9 % des élèves sont "à l'heure", 3,4 % "en avance", et 39,7 % "en retard". Encore est-il nécessaire, avant de tirer des conclusions de ce constat, de savoir ce que signifie ici l'âge : par exemple, s'agit-il de l'âge à la rentrée scolaire de 1997, au premier janvier 1998, ou à la fin de l'année scolaire ? L'ouvrage d'où ce tableau est extrait nous apprend que c'est la seconde hypothèse qui est la bonne : "l'âge est le nombre d'années révolues au 1er janvier inclus dans l'année scolaire" [2].

La variable "âge" présente toutes les propriétés des variables nominales et ordinales : les catégories d'âge sont bien exclusives les unes des autres, et exhaustives pour la population considérée (à chaque élève correspond une classe d'âge et une seule) ; ces classes d'âge s'ordonnent selon une dimension unique, et l'on ne peut modifier cet ordre sans altérer profondément la signification de la variable. Cette variable présente en outre une caractéristique supplémentaire : elle possède une *unité de mesure* (ici : l'année), permettant d'évaluer les intervalles entre deux quelconques de ces catégories. C'est une variable *mesurable* (ou : *métrique*). Dans les sciences sociales, le revenu individuel ou par ménage, le nombre de personnes par foyer, la dimension de l'agglomération de résidence, sont des variables mesurables.

L'intérêt des variables mesurables est qu'il est possible d'effectuer des opérations arithmétiques sur les valeurs prises par ces variables. On peut par exemple calculer une *moyenne* de la distribution : l'âge moyen des élèves de troisième est égal à la somme des âges, divisée par le nombre des élèves. À partir des données dont nous disposons, le calcul ne peut être qu'approximatif, et ceci pour deux raisons : 1) les âges ont été regroupés par années, mettant dans la même catégorie des enfants ayant jusqu'à 364 jours de différence, et distinguant éventuellement des enfants dont la date de naissance ne diffère que d'un seul jour (31 décembre, 1er janvier) ; 2) les catégories extrêmes ne sont bornées que d'un seul côté (13 ans *et moins*, 16 ans *et plus*), et l'on ne dispose donc pas de la répartition réelle des âges (mêmes regroupés) dans sa totalité. Ces simplifications ont évidemment été adoptées pour rendre le tableau plus lisible, mais en contrepartie elles diminuent la précision des calculs que l'on peut faire sur cette distribution.

Si l'on avait voulu calculer un âge moyen (au 1er janvier 1998) avec le maximum de précision, il aurait fallu disposer de l'âge de chaque élève exprimé en jours (ce qui est possible, connaissant sa date de naissance). Toutefois, si l'on se contente d'un *ordre de grandeur*, et non d'une mesure précise au jour près, on peut calculer un âge moyen à partir des valeurs dont nous disposons : on obtient 14,45 ans, soit 14 ans 5 mois et 15 jours *approximativement* (le 165ième jour de l'année est en fait le 13 juin, et non le 15, en année non bissextile). Le schéma 2 (qui traduit graphiquement la colonne "effectifs" du tableau 7) permet de voir que cette valeur doit être utilisée avec réserve, puisqu'il est probable qu'à

[1] *Op. cit.*, page 86.
[2] *Op. cit.*, page 60.

l'intérieur de chaque classe d'âge, la répartition des âges "réels" (calculés en jours) n'est pas symétrique, comme le suggère la courbe ajustée à l'histogramme [1].

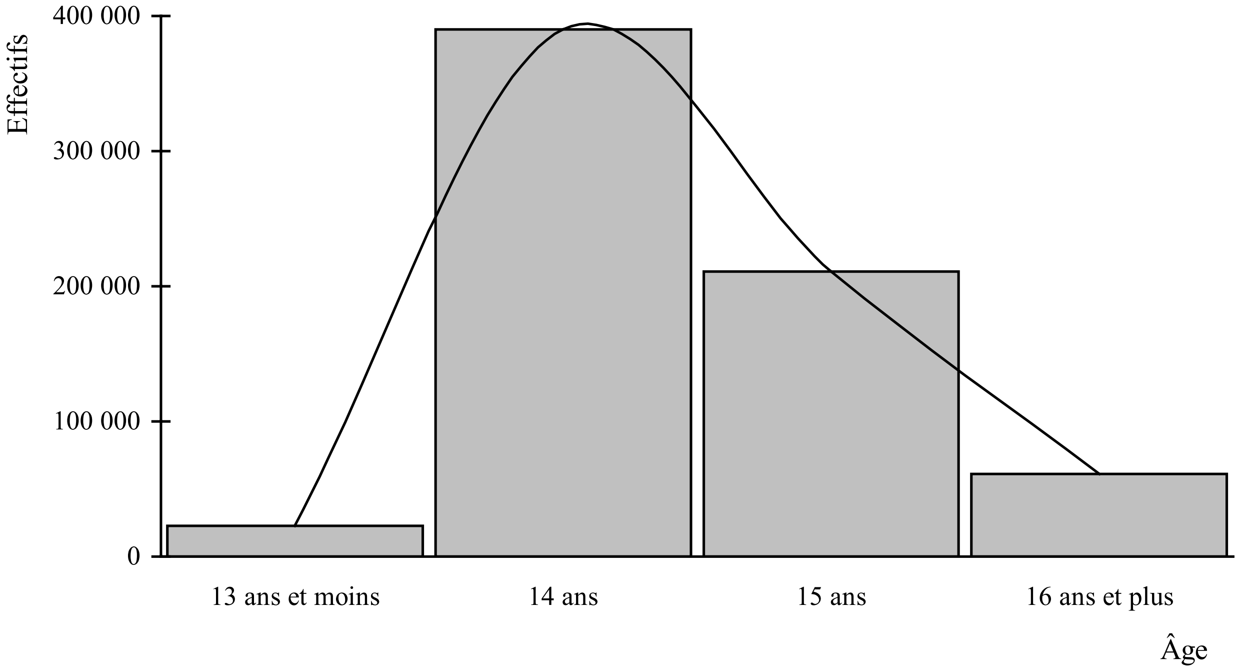


Schéma 2. Répartition par âge des élèves de troisième générale en 1997-1998.

En supposant que l'âge moyen ainsi calculé soit acceptable dans son principe, comment l'interpréter ? Affirmer que l'âge moyen au 1er janvier 1998 des élèves de troisième générale est de 14 ans 5 mois et 15 jours serait cependant une erreur. Pour s'en assurer, il suffit de revenir à la définition de l'âge dans cette recherche : puisqu'il s'agit de l'âge *révolu* au 1er janvier 1998, les élèves classés dans la catégorie "14 ans" ont en réalité entre 14 ans (s'ils sont nés le 31 décembre 1997) et 14 ans 11 mois et 30 jours (s'ils sont nés le 1er janvier 1997). Il faut donc ajouter 6 mois à l'âge moyen calculé précédemment pour avoir une meilleure approximation de l'âge moyen réel. Encore cette approximation n'échappe-t-elle pas à toute critique, puisque l'ajout d'exactement 6 mois au résultat du premier calcul suppose que la classe "14 ans" a un âge moyen "réel" (calculé au jours près) de 14 ans 6 mois, ce que nous ignorons ; nous pouvons toutefois raisonnablement supposer que cette hypothèse n'est pas trop éloignée de la réalité, et conclure que l'âge moyen des élèves de troisième générale au 1er janvier 1998 est de l'ordre de 14 ans 11 mois et 15 jours. Le même raisonnement doit évidemment être appliqué au mode et à la médiane de la distribution.

[1] L'ajustement de cette courbe n'est qu'approximatif. Elle a l'intérêt d'indiquer l'allure générale que prendrait une distribution continue basée sur l'âge "vrai" (calculé à partir des âges bruts).

1.4. Récapitulation et compléments.

Nous revenons dans ce sous-chapitre sur une notion théorique de grande importance en statistiques descriptives, celle des *niveaux de mesure* (ou types de variables) ; c'est en effet selon le niveau des variables que l'on choisira les outils statistiques qui conviennent. Nous abordons ensuite le problème pratique du traitement des "non réponses". Nous terminons sur quelques éléments relatifs à la théorie des sondages.

a) Les niveaux de mesure.

Les trois exemples de distribution à une seule variable que nous avons présentés illustrent les trois niveaux de mesure les plus couramment utilisés dans les sciences sociales. Chaque niveau est caractérisé par un ensemble d'hypothèses relatives aux propriétés de la variable. Le terme *mesure* est utilisé ici dans son sens le plus large, désignant toute opération qui fait correspondre une valeur numérique à une observation ou un ensemble d'observations (par exemple, une fréquence à une PCS). L'expression *niveaux de mesure* exprime l'idée d'un ordre, d'une hiérarchie sur les propriétés attribuées aux variables, allant du niveau le plus faible (utilisant peu d'hypothèses) au niveau le plus puissant (le plus riche en propriétés).

Les hypothèses les plus simples sont celles qui permettent de classer les "objets" étudiés (individus, groupes, entreprises, communes, etc.), et de leur donner un *nom* collectif ; c'est pourquoi le niveau de mesure qui leur correspondent est appelé *niveau nominal*. Prenons une population décrite par une variable nominale, par exemple la profession de ses membres ; chaque individu sera caractérisé (identifié, "nommé") par sa profession. Si l'on compare deux individus quelconques de cette population, on ne peut dire qu'une seule chose à leur sujet : ils ont la même profession, ou ils sont de professions différentes. Le niveau nominal est fondé sur la relation d'*identité* (relation réflexive, symétrique, transitive) : deux individus sont soit identiques (ils ont la même profession), soit différents du point de vue de la variable nominale considérée. Cette formalisation entraîne deux conséquences : 1) tout individu appartient à une catégorie (quitte à créer la catégorie des "sans profession") ; 2) un individu ne peut appartenir qu'à une seule catégorie (s'il exerce simultanément deux professions, il faudra soit créer une catégorie spéciale pour ce cas particulier, soit décider à laquelle des deux catégories possibles il sera rattaché).

Sans modifier les règles d'affectation d'un individu à une catégorie et la relation d'identité qui les sous-tend, on peut ajouter une relation *d'ordre strict* entre les catégories elles-mêmes (relation anti-réflexive, anti-symétrique, transitive). On considèrera par exemple qu'une profession est plus qualifiée, plus prestigieuse, ou mieux rétribuée, qu'une autre. Le niveau *ordinal* permet par conséquent de comparer et de classer les catégories entre elles, et de décider si, par rapport à la variable ordinale considérée, l'une d'elles est supérieure, égale (identique), ou inférieure à l'autre. Pour que cette comparaison puisse être faite entre toutes les catégories, il faut que la relation d'ordre s'applique à la totalité de celles-ci (ordre strict *total*). Dans la plupart des travaux sur la mobilité sociale, la profession de la personne interrogée et celle de son père servent d'indicateurs de la position et de l'origine sociales ; c'est le fait de considérer la profession comme une variable ordinale qui permet de parler de "mobiles ascendants" et de "mobiles descendants".

Aux propriétés des variables ordinales, on peut ajouter une propriété supplémentaire particulièrement puissante, celle de la "mesurabilité" au sens strict. Parmi les exemples les plus courants de variable *mesurable* (métrique) on peut citer l'âge, le revenu, la dimension de la famille, la taille de l'agglomération de résidence. Dans ce type de variables, on ajoute à la relation d'ordre strict la définition d'*intervalles* (ou de distances) entre les catégories (et par conséquent entre les individus), intervalles mesurés à l'aide d'une *unité* (année, Francs, nombre de personnes) ; il est alors possible de dire *de combien* une personne ou une collectivité est "supérieure" (plus qualifiée, plus nombreuse) qu'une autre. Le grand intérêt de ce type de variables est qu'elles autorisent tous les calculs arithmétiques usuels (comme le calcul de moyennes). Comparer les professions selon leurs revenus moyens revient à considérer la PCS comme une variable mesurable ; c'est ce que l'on fait lorsque l'on dit par exemple que l'éventail des salaires varie du simple au double entre telle profession et telle autre.

L'exemple de la PCS permet de comprendre que les niveaux de mesure sont des *modèles*, autrement dit des constructions abstraites indispensables au chercheur en sciences sociales pour décrire et analyser ses données ; c'est ce que signifie l'affirmation : "les variables sont *construites*". Le choix du niveau de mesure adapté à une variable est une décision qui appartient en propre au sociologue, car lui seul a (de par sa formation théorique et son expérience du terrain) une connaissance intime des données, et donc la compétence nécessaire pour prendre cette décision. Autant il est nécessaire de faire appel aux mathématiciens ou aux statisticiens pour comprendre les propriétés formelles de ces modèles, autant il serait imprudent de laisser à ceux-ci la responsabilité d'une décision d'ordre purement sociologique.

Ces trois niveaux de mesure ne sont pas les seuls possibles : les variables sur lesquelles le sociologue travaille habituellement ont souvent des propriétés plus complexes que celles sur lesquelles sont construits ces modèles. Par exemple, nous avons vu que, lors même que l'on considère la PCS comme une variable nominale, on admet implicitement que les catégories socioprofessionnelles sont plus ou moins proches les unes des autres ; sinon, il n'y aurait aucune légitimité à procéder à des regroupements. Le seul avantage de ces modèles fortement réducteurs est qu'à chacun d'eux correspond un arsenal d'outils statistiques utiles pour une analyse plus poussée des données. L'alternative laissée au chercheur est soit de se rabattre sur un modèle appauvrissant, mais commode, soit de vouloir conserver à tout prix la richesse des informations dont il dispose, et renoncer à les exploiter.

b) Les opérations possibles sur les tableaux à une dimension.

Les hypothèses faites sur la nature formelle des variables ont pour but de faciliter la description, et donc la compréhension, du contenu des tableaux. Toute description est nécessairement réductrice ; les mots du langage, et à plus forte raison les concepts qu'ils expriment, ne retiennent de la réalité que quelques traits : lorsque le sociologue parle des ouvriers ou des cadres supérieurs, ils ne retient des personnes qu'il désigne ainsi que certains traits qu'ils ont en commun. La description statistique pousse un peu plus loin l'abstraction. La meilleure description d'une distribution, du point de vue de la fidélité et de l'exhaustivité, serait la distribution elle-même. En pratique, on extrait de cette distribution une information particulière dont on a besoin dans le cours de la description du tableau. Les questions que l'on se pose le plus fréquemment à propos d'une distribution sont de trois types : 1) quelle part occupe dans la distribution telle catégorie de la population ? 2) quelle est la caractéristique dominante de la distribution ? 3) dans quelle mesure la population est-elle homogène du point de vue de la variable considérée ?

1) Le premier type correspond aux questions de la forme : "quelle est la proportion d'élèves *en avance* (de moins de 14 ans) dans les classes de troisième ?", ou encore : "que représentent les agriculteurs dans l'ensemble de la population active ?". Pour y répondre, on peut calculer les fréquences relatives (variant de 0 à 1), ou plus couramment les pourcentages (variant de 0 à 100). Aux niveaux ordinal et métrique, on utilisera si nécessaire les pourcentages cumulés.

2) Le second type de questions porte sur la *valeur centrale* d'une distribution : "la population active française est-elle composée surtout de travailleurs de l'industrie ou des services ?" ; ou encore : "quel est en gros l'âge 'normal' des élèves de troisième ?". Selon les propriétés attribuées à la variable, on recourra pour y répondre au mode, à la médiane, ou à la moyenne de la distribution.

3) L'hétérogénéité d'une distribution est l'objet des *mesures de dispersion*. Au niveau nominal, une distribution est d'autant plus homogène que plus d'individus possèdent la caractéristique dominante (que le *mode* de la distribution est proportionnellement plus important), ou que la majorité de la population se concentre sur un petit nombre de catégories ; le maximum d'hétérogénéité est atteint lorsque toutes les catégories ont le même effectif (distribution rectangulaire, ou équipartition). Dans le premier cas, le coefficient d'hétérogénéité sera proche de *zéro* ; dans le second, il prendra sa valeur maximum. À ce niveau, l'un des coefficients de dispersion les plus utilisés est l'*entropie*. Au niveau *métrique*, la dispersion est d'autant plus grande que les individus s'écartent plus de la moyenne. Elle est maximum lorsque les individus se répartissent également entre les deux valeurs extrêmes ; elle est nulle lorsque tous les individus se trouvent dans la moyenne. Le coefficient le plus courant à ce niveau est la *variance* (et sa racine carrée, l'*écart-type*). Au niveau ordinal, les mesures de dispersion dont on dispose (écart interquartile, par exemple) ne sont guère utilisées en sociologie, car elles n'ont pas la qualité des coefficients existant pour les autres niveaux de mesure ; faute de coefficient mieux adapté, on peut toujours se rabattre sur l'entropie.

Les valeurs prises par ces indices ou coefficients n'ont pas toujours beaucoup d'intérêt en elles-mêmes ; l'examen visuel de la distribution suffit le plus souvent pour repérer (au moins approximativement) sa valeur centrale, ou pour estimer son degré d'hétérogénéité. L'utilité de ces mesures statistiques est de permettre des comparaisons rigoureuses entre différentes distributions. Lorsque, par exemple, on dit que la population d'un pays est plus jeune que celle d'un autre pays, c'est que l'on a constaté que l'âge moyen dans le premier est inférieur à l'âge moyen dans le second ; on peut en outre préciser l'ampleur de la différence, ce qu'un simple examen visuel de la distribution ne permet évidemment pas.

Question	Variable nominale	Variable ordinale	Variable mesurable
Quelle part occupe dans la distribution telle catégorie (ou tel ensemble de catégories) ?	fréquences relatives brutes, pourcentages bruts	fréquences relatives brutes ou cumulées, pourcentages bruts ou cumulés	
Quelle est la caractéristique dominante de la distribution (*valeur centrale*) ?	mode	mode, médiane	mode, médiane, moyenne
Dans quelle mesure la population est-elle homogène (*dispersion*) ?	entropie	écart interquartile	variance (et écart-type)

Indices statistiques les plus usuels pour décrire une distribution à une variable.

D'autres questions également importantes se posent à propos d'une distribution unidimensionnelle, en particulier celles relatives à la *forme* de la distribution. Par exemple, est-on devant une courbe "en cloche", "en U", "en J" ou "en L" ? À chaque forme de courbe peut être ajusté un modèle mathématique particulier, avec des hypothèse spécifiques ; par exemple, la distribution de Laplace-Gauss ("courbe en cloche") correspond à l'influence, sur la variable considérée, d'un grand nombre de variables indépendantes les unes des autres, et dont par conséquent les effets s'ajoutent. Certains politologues pensent que la courbe "en cloche" caractérise les "opinions privées", comme celles que l'on déclare en tête à tête à un interlocuteur (qui peut être un enquêteur), et la courbe "en U" les "opinions publiques", que l'on affiche par opposition à une opinion contraire (comme les opinions que l'on peut recueillir dans les groupes de discussion).

Si la distribution apparaît (au moins vaguement) comme une "courbe en cloche", est-elle unimodale (a-t-elle une seule "bosse") ? Si la distribution a par exemple deux "bosses" (distribution bimodale), on peut en conclure qu'il doit exister en réalité deux sous-populations hétérogènes (selon la variable considérée) qui se trouvent mélangées. La recherche pourra se fixer comme objectif ultérieur d'identifier les variables qui permettent de distinguer ces deux populations ; en attendant, il va de soi que, même si l'on peut toujours repérer une médiane ou une calculer moyenne sur une telle distribution, cet indicateur de valeur centrale n'aura aucune utilité, puisque dépourvu de signification sociologique. La forme de la distribution a de plus une grande importance si l'on veut ajuster un modèle théorique (loi de Gauss, loi binomiale, etc.) à celle-ci, afin de pousser plus loin l'interprétation.

Une autre question concerne la *symétrie* de la distribution, que celle-ci soit unimodale ou en U. Il existe d'ailleurs des coefficients d'asymétrie, mais ceux-ci sont assez peu utilisés par les sociologues. Il est vrai qu'un simple examen visuel suffit pour établir si la distribution est ou non unimodale, et si elle est ou non symétrique. Cet examen doit être fait systématiquement, préalablement à tout calcul de valeur centrale ou de dispersion, afin de préparer l'interprétation sociologique des valeurs prises par ces coefficients.

c) Le problème des non réponses.

Que faire lorsque les données dont on dispose comportent des "non réponses" ? Le problème auquel on est ici confronté tient à l'hétérogénéité de cette catégorie : les personnes interrogées ont-elles refusé de répondre à la question posée parce qu'elles trouvaient la question indiscrète, parce qu'elles n'étaient pas d'accord avec la manière dont la question était formulée, parce qu'elles ne se sentaient pas concernées par le problème évoqué, parce qu'elles estimaient avoir déjà répondu à une question analogue, parce qu'elles n'avaient réellement pas d'avis sur la question (cela ne les avait jamais intéressées), parce qu'elles auraient souhaité pouvoir réfléchir plus longuement à la question avant de donner une réponse, ou encore parce qu'elles ne comprenaient pas les termes de la question ou la jugeaient trop difficile pour eux ?

Lorsque l'on considère que la variable sur laquelle on travaille est une variable nominale, conserver telle quelle la catégorie des "non réponses" ne pose pas de problème formel. Seule subsiste la difficulté d'interprétation de cette catégorie ; c'était le cas pour la PCS "chômeurs n'ayant jamais travaillé" du tableau 1, puisque, contrairement aux retraités par exemple, ils n'étaient pas classables selon une profession. En revanche, si l'on attribue à la variable des propriétés ordinales, ou des propriétés métriques, cette catégorie n'a pas sa place dans une variable de ce type. Comment en effet repérer une médiane ou calculer une moyenne dans ces conditions ?

La solution que l'on adopte généralement n'est guère satisfaisante ; mais on n'en connaît pas de meilleure pour le moment : on choisit habituellement de négliger les "non réponses" dans le traitement de la variable, comme nous l'avons fait dans le tableau 5. Une justification possible de cette manière de faire serait que les "non réponses" se répartissent aléatoirement au sein des personnes interrogées. En d'autres termes, les personnes qui n'ont pas répondu à une question donnée ne se distingueraient pas (selon les variables sociologiques usuelles, ou selon d'autres variables, psychologiques, contextuelles, etc.) de celles qui y ont répondu ; on pourrait alors admettre que, si elles avaient répondu, leurs réponses se répartiraient comme celles des personnes qui ont répondu. Il serait alors légitime de répartir les "non répondants" dans les autres catégories de réponses, proportionnellement à l'effectif de ces catégories. On vérifiera qu'en procédant ainsi, on obtient les mêmes pourcentages que ceux calculés sur les seuls répondants.

Des études empiriques ont montré que les "non répondants" avaient en réalité des caractéristiques sociologiques bien particulières (en termes de niveau économique et culturel, par exemple), et que par conséquent les "non réponses" étaient des réponses chargées de sens [1]. Il est donc important de les prendre en compte dans une première phase de l'analyse des réponses. Si ces "non réponses" ne sont pas trop nombreuses, on peut ensuite, à condition de le signaler explicitement, les éliminer de la distribution, pour pouvoir utiliser les propriétés formelles de la variable. On gardera toutefois à l'esprit, lors de l'interprétation de ces données, que celle-ci ne porte que sur une partie des personnes interrogées.

[1] Voir par exemple : Guy Michelat, et Michel Simon, "Les 'sans réponse' aux questions politiques", *Pouvoirs*, 33 (avril 1985), pages 41-56.

d) Population et échantillon.

Lorsque l'on ne dispose pas de sources administratives, et que l'on désire connaître la fréquence de certaines caractéristiques dans une population (sexe, profession, âge, revenus, etc.), la meilleure démarche consiste à recueillir ces informations en interrogeant chaque individu de la population ; c'est ce que l'on fait dans les recensements. Il arrive fréquemment que, pour des raisons de délais ou de coût par exemple, il soit impossible d'interroger la totalité de la population étudiée. Peut-on alors connaître avec exactitude la fréquence cherchée en n'interrogeant qu'une partie des individus ? Évidemment non.

Par contre, si l'on accepte une certaine imprécision dans les informations recherchées, il est possible, à certaines conditions, d'obtenir une valeur approchée des fréquences désirées (ceci vaut d'ailleurs pour toute autre donnée statistique, comme la moyenne).

Ces conditions font l'objet de la *théorie des sondages* [1]. Les principes de base de cette théorie, dont la compréhension est indispensable aux sociologues, sont les suivants :

1) La sous-population interrogée (l'*échantillon*) doit être sélectionnée de telle sorte qu'elle constitue un modèle réduit de la population visée. Pour obtenir ce résultat, la procédure idéale est la sélection *aléatoire* (tirage au hasard) dans une *base de sondage* (liste, sans omission ni répétition, de tous les individus de la population). En pratique, diverses autres manières de constituer un échantillon *représentatif* d'une population donnée sont utilisées ; mais l'échantillonnage aléatoire demeure la procédure de référence, dont les autres procédures ne sont (au mieux) que des approximations.

2) Si l'échantillon est véritablement représentatif, il est possible de calculer les *limites de confiance* entre lesquelles la valeur cherchée a une probabilité de se trouver. Le risque de se tromper (la probabilité d'erreur) accepté usuellement par les sociologues est de 5 % (seuil de P = 0,05). Cela signifie qu'ils espèrent que les limites de confiance seront conformes à la réalité dans 95 % des cas en moyenne, et que la valeur cherchée (la valeur "vraie") a 95 chances sur 100 de se trouver entre ces limites.

3) Plus le risque d'erreur que l'on accepte est faible, moins l'estimation est précise. Si l'on désire par exemple que les limites de confiance soient conformes à la réalité dans 99 % des cas en moyenne (risque d'erreur de 1 % seulement, P = 0,01), il faudra élargir la "fourchette", de sorte que la valeur cherchée ait 99 chances sur 100 de se trouver entre les limites de confiance.

4) Pour un seuil de probabilité donné, la précision est d'autant meilleure que la taille de l'échantillon est plus grande. À titre indicatif, pour doubler la précision d'une estimation (toutes choses égales par ailleurs), il faut quadrupler la taille de l'échantillon (la précision croît comme la racine carrée de la taille de l'échantillon).

5) Toutes choses égales par ailleurs, la précision de l'estimation est meilleure pour les fréquences très petites ou très grandes (on le vérifiera sur les tableaux 4 et 6). Lorsque les

[1] Pour s'initier aux problèmes liés aux sondages, on lira : Anne-Marie Dussaix, Jean-Marie Grosbras, *Les sondages : principes et méthodes*, Paris, PUF (*Que sais-je ?* n° 701), 1993.

politologues tentent de prévoir l'issue d'un référendum, ou d'un scrutin à deux candidats, à partir d'un échantillon des premiers dépouillements, la prévision est d'autant plus difficile que le scrutin est plus serré (que les estimations sont plus proches de 50 %).

Chapitre 2. Distributions à deux variables.

Les distributions à deux variables sont également appelées *tris croisés*, ou *tableaux croisés* à deux dimensions, ou encore *tableaux à double entrée*. Il y a distribution à deux variables lorsque l'on dispose du dénombrement des observations pour chaque couple de valeurs prises par les deux variables considérées.

2.1. Premier exemple : taux de suicide des hommes célibataires selon l'âge (France, 1889-1891).

Le tableau 8 est extrait du tableau XXI du *Suicide*, d'Émile Durkheim [1]. Il présente l'apparence d'une distribution à une seule variable, puisqu'il n'a qu'une colonne. Il ne s'agit ici que d'un artifice de présentation, rendu possible parce que l'une des deux variables (le suicide) est dichotomique (elle ne peut prendre que deux valeurs : oui ou non).

Le tableau d'où il est tiré présente les "suicides commis par 1 000 000 d'habitants de chaque groupe d'âge et d'état civil, année moyenne". La population décrite dans le tableau 8 est l'ensemble des Français célibataires de sexe masculin, vivant en 1889, 1890 ou 1891. Les deux variables considérées sont l'âge, variable mesurable dont les valeurs ont été regroupées en neuf catégories, et le comportement suicidaire. Comme la proportion de non suicidaires se déduit automatiquement du taux de suicides, l'auteur n'a pas jugé utile de la faire figurer dans le tableau.

Âge	Suicides pour un million de célibataires de chaque groupe d'âge
15-20 ans	113
20-25 ans	237
25-30 ans	394
30-40 ans	627
40-50 ans	975
50-60 ans	1 434
60-70 ans	1 768
70-80 ans	1 983
Au-delà	1 571

Tableau 8. Taux de suicide moyen selon l'âge (hommes célibataires, France 1889-91).

Ce tableau a été établi à partir de l'ensemble des procès verbaux dressés à l'occasion d'un suicide en France entre 1889 et 1891. Chaque procès verbal comportait diverses informations : date et heure du suicide ; date de naissance, sexe et état civil du suicidé ; motifs

[1] Émile Durkheim, *Le suicide*, Paris, Alcan,1897, (rééditions PUF), page 183.

présumés du suicide ; moyens utilisés pour mettre fin à ses jours ; etc. Tels quels, ces procès verbaux représentaient un volume d'informations beaucoup trop important pour pouvoir être assimilé dans sa totalité. Pour analyser toutes ces informations, il a donc fallu au préalable les *condenser*, c'est à dire les sélectionner et les simplifier.

À partir de ses hypothèses, l'auteur a décidé de *sélectionner* les variables qu'il supposait le plus liées au suicide : le sexe, l'âge, la situation de famille. Afin de les traiter plus facilement, il lui a fallu *simplifier* ces données, en regroupant des cas individuels différents certes, mais voisins : en ne distinguant que neuf catégories d'âge, et trois catégories de situations de famille.

Cette manière de procéder néglige une grande partie de l'information contenue dans les procès verbaux ; elle réduit l'infinie diversité des situations individuelles à un petit nombre de types normalisés. Les renseignements rassemblés sont ainsi devenus assimilables, mais au prix d'une réduction qui les appauvrit considérablement. Cette démarche est non seulement légitime : elle est inévitable, même au niveau le plus descriptif, le plus proche des données. Toute tentative pour décrire des "faits sociaux" impose un effort d'abstraction, de sélection des informations que l'on juge "intéressantes", ou "pertinentes" ; même le recueil des données suppose un commencement d'*interprétation* qui détermine le choix de ce que l'on observe, et de la manière dont on le classe dans des catégories. Cette démarche n'est pas le propre des méthodes "quantitatives" : les méthodes d'analyse de contenu ou d'observation ne procèdent pas autrement.

a) Présentation du tableau sous une forme complète.

Le tableau 8 peut être mis sous une forme moins condensée, faisant apparaître les taux de non suicides. Dans la présentation ci-dessous (tableau 9), les taux de suicides ont été ramenés à des "pour mille" (‰), plus usuels que les taux utilisés par Durkheim. En outre, on a légèrement modifié l'intitulé des classes d'âge, en supposant que les "15-20 ans" incluaient bien les célibataires âgés de 20 ans, ce qui implique que la classe suivante commence à 21 ans.

Âge	Suicides	Non suicides	Base : 1 000 ‰ =
15-20 ans	0,113‰	999,887‰	?
21-25 ans	0,237‰	999,763‰	?
26-30 ans	0,394‰	999,606‰	?
31-40 ans	0,627‰	999,373‰	?
41-50 ans	0,975‰	999,025‰	?
51-60 ans	1,434‰	998,566‰	?
61-70 ans	1,768‰	998,232‰	?
71-80 ans	1,983‰	998,017‰	?
Plus de 80 ans	1,571‰	998,429‰	?
Ensemble	? ‰	? ‰	?

Tableau 9. Taux de suicide moyen selon l'âge (hommes célibataires, France 1889-91).

Cette nouvelle présentation met en évidence les lacunes du tableau d'origine : il manque les *bases* des proportions, c'est à dire les effectifs de chaque classe d'âges, effectifs à partir desquels ces taux ont été calculés. Inutiles pour le propos de l'auteur, ces effectifs auraient pu être utilisés dans l'analyse de ce tableau : connaissant l'effectif de chaque classe d'âge, et donc de l'ensemble des célibataires étudiés, il aurait été possible par exemple de calculer un âge moyen au suicide, et de connaître le taux moyen de suicides pour les célibataires. En outre, ces effectifs nous auraient permis de regrouper certaines classes d'âges en vue de comparer ce tableau à d'autres tableaux de même nature, mais utilisant un découpage différent.

b) Interprétation du tableau.

Tel qu'il se présente, ce tableau nous renseigne utilement sur les effets de l'âge sur le taux de suicides des célibataires : ce taux croît régulièrement avec l'âge jusqu'à 80 ans ; au delà, on observe une diminution difficile à interpréter. Il est possible que le faible effectif de cette dernière classe d'âge rende les taux s'y rapportant plus sensibles aux fluctuations annuelles.

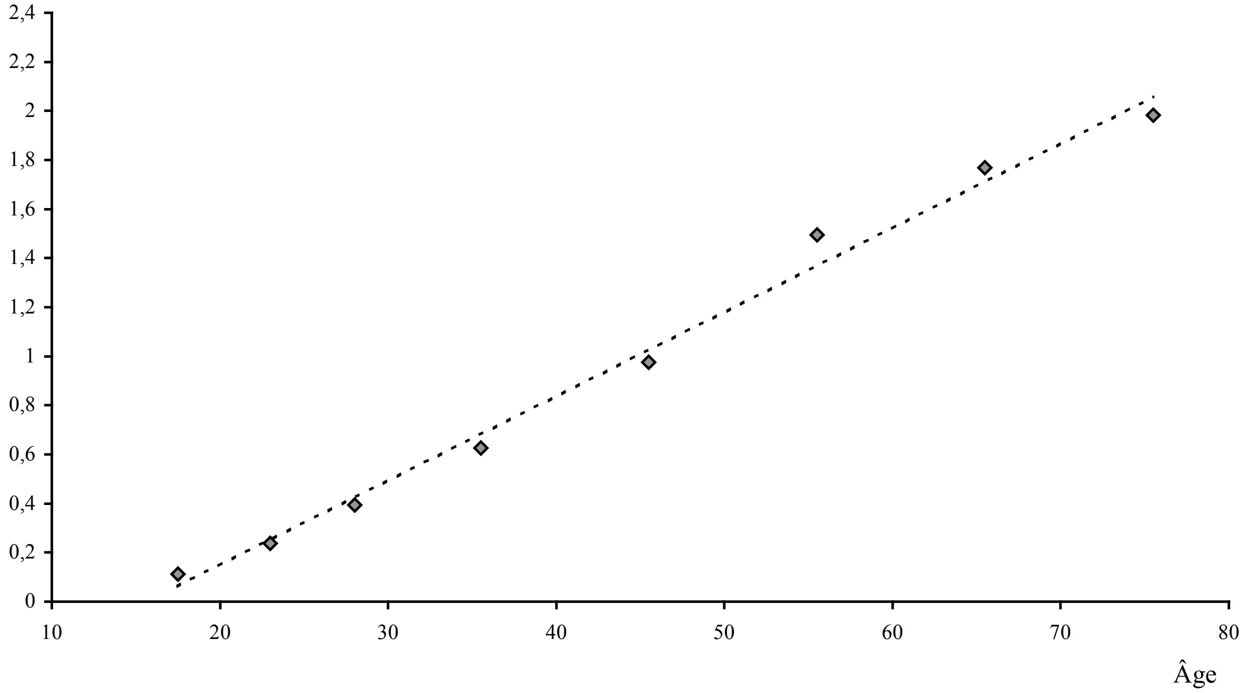


Schéma 3. Taux de suicides selon l'âge (célibataires, France, 1889-91).

Une lecture trop rapide de ce tableau risque de passer à côté d'un fait pourtant remarquable : la progression du taux de suicides selon l'âge est non seulement régulière (*monotone*) jusqu'à 80 ans, elle est aussi quasiment *linéaire*. Cette observation est masquée par le choix des classes d'âges fait par l'auteur : leur dimension, de 5 années à l'origine, passe à 10 à partir de 30 ans. Si l'on prend la peine de représenter les variations de ce taux sur un graphique, en prenant l'âge médian pour représenter chaque classe d'âges, et en négligeant la dernière classe (qui ne permet pas de déterminer un âge médian, parce que non bornée), on obtient presque une droite (schéma 3) [1]. Cet exemple illustre l'une des fonctions principales des tableaux à double entrée : décrire les variations concomitantes des valeurs prises simultanément par deux variables.

Avant de conclure sur la lecture de ce tableau, ajoutons que, pour bien l'interpréter, il faudrait connaître exactement la manière dont il a été construit : les populations prises comme référence pour chaque année considérée sont-elles les célibataires vivant au 1er janvier, ou à une autre date ? Les statistiques portant sur trois années, comment a-t-on fait pour calculer un taux moyen de suicides par classe d'âges ? Enfin, dispose-t-on par ailleurs d'informations permettant de distinguer l'effet de l'âge de l'effet de génération (influence de l'histoire : vague suicidaire liée à une crise économique, par exemple) ?

---

[1] La droite figurant sur ce schéma a été ajustée au plus près des points correspondants aux taux de suicides, selon la méthode "des moindres carrés" (telle que la somme des carrés des distances des points à cette droite soit minimale).

2.2. Deuxième exemple : l'origine socioprofessionnelle des étudiants français en 1997-98.

Ce tableau présente les effectifs d'étudiants inscrits dans l'enseignement supérieur en 1997-98, selon la catégorie socioprofessionnelle des parents et la filière choisie [1]. Cette statistique officielle a été établie à partir des dossiers d'inscription des étudiants dans les universités françaises (métropole seulement) à la rentrée 1997. Les étudiants étrangers n'y figurent pas (ils font l'objet d'un autre tableau), ni les Français étudiant à l'étranger. Il est bon de garder ce dernier point à l'esprit au moment de l'interprétation du tableau, puisqu'il est possible que les études à l'étranger soient plus fréquentes dans les milieux plus aisés, ou plus instruits.

Le tableau 10 présente les effectifs des étudiants des deux sexes et des trois cycles d'études, selon la filière suivie et la catégorie socioprofessionnelle du chef de famille. Les auteurs distinguent les IUT des filières universitaires proprement dites, et classent ces filières par disciplines. On notera le regroupement des sciences et techniques des activités physiques et sportives (STAPS) avec les autres sciences, et l'assimilation des sciences humaines aux lettres. La catégorie socioprofessionnelle (PCS) du chef de famille utilise la nomenclature de l'INSEE de 1982 ; elle sert ici d'indicateur de l'origine sociale des étudiants.

a) Lecture des marges du tableau.

À la rentrée 1997, il y avait en France métropolitaine 1 303 177 étudiants de nationalité française inscrits dans les universités. Les élèves des grandes écoles et des classes préparatoires, comme d'ailleurs ceux des sections de techniciens supérieurs, ne sont pas compris dans ce tableau. Les totaux de colonnes nous renseignent sur la répartition globale des effectifs par filières : les lettres, fortement majoritaires, représentent plus du tiers des étudiants (35,1 %), suivies des sciences (23,7 %) ; les quatre filières restantes représentent chacune à peu près un dixième des étudiants : 12,9 % en droit, 10,3 % en économie, 9,7 % dans les disciplines de santé, et 8,4 % en IUT.

Il faut se garder d'en conclure par exemple que 35 % des étudiants ont *choisi* les études "littéraires", tandis que 8 % seulement se sont dirigés vers les IUT. En premier lieu, une telle affirmation ne tiendrait pas compte de possibles réorientations après le choix initial. Mais surtout, ce tableau ne porte pas sur les *choix* : il décrit l'ensemble des étudiants *inscrits* dans les universités, et non les étudiants inscrits en première année d'enseignement supérieur. Même si la durée des études était identique dans chacune des filières, pour induire de ce tableau les *choix* des étudiants, il faudrait supposer que les taux d'abandons en cours d'études sont du même ordre de grandeur dans toutes les filières. Comme la durée des cursus varie de deux années pour les IUT à deux, quatre, voire sept années ou plus dans la plupart des autres

[1] Tableau 1, page 167 de *Repères & références statistiques sur les enseignements et la formation*, Ministère de l'Éducation nationale, de la Recherche et de la Technologie, Paris, 1998.

filières, on commettrait une erreur grave en décrivant cette marge du tableau en termes de choix de filière (en disant par exemple : "les IUT *accueillent* 8,4 % des étudiants", ou encore : "9,7 % *se dirigent vers* les disciplines de santé"). On mesure l'ampleur de l'erreur ainsi commise selon les filières en comparant ces pourcentages à ceux enregistrés pour les nouveaux entrants [1].

Les totaux de ligne décrivent la population étudiante (universités seulement) selon la profession du chef de famille. On observe que les milieux "aisés" sont fortement surreprésentés : près des deux tiers des étudiants sont issus de familles dont le chef de ménage est cadre supérieur ou de profession libérale (32,4 %), contre seulement 11,2 % qui ont un père ouvrier et 11,9 % un père employé, alors que ces catégories sont numériquement les plus importantes dans l'ensemble de la population.

Origine socioprofessionnelle (PCS)	Droit	Économie	Lettres	Sciences et STAPS	Santé	IUT	Ensemble Université
Agriculteurs	3 094	3 337	9 914	8 090	2 416	4 034	30 885
Artisans, comm., chefs d'entreprise	15 361	12 118	32 793	21 382	7 969	9 470	99 093
Prof. libérales, cadres supérieurs	61 790	41 104	126 512	109 537	55 001	27 676	421 620
Professions intermédiaires	23 898	20 472	81 448	56 817	15 723	21 913	220 271
Employés	20 499	16 991	59 669	34 316	7 459	15 727	154 661
Ouvriers	15 448	16 848	56 011	32 756	6 459	18 372	145 894
Retraités, inactifs	14 472	12 576	43 769	20 194	6 183	6 945	104 139
Indéterminé	13 432	10 232	47 305	25 245	24 787	4 613	125 614
Ensemble	167 994	133 678	457 421	308 337	125 997	108 750	1 302 177

Tableau 10. Origine socioprofessionnelle des étudiants français dans les principales filières de l'enseignement supérieur en 1997-1998 (France métropolitaine).

Il est tentant de comparer cette distribution de la population étudiante avec celle de l'ensemble de la population, afin d'en induire les "chances" d'accéder à l'enseignement supérieur selon le milieu social d'origine. On peut toujours calculer pour chaque PCS le rapport entre le nombre d'étudiants issus de cette catégorie et l'effectif à la même date de cette catégorie ; les différences d'accès aux études supérieures sont tellement grandes qu'un tel rapport suffit pour en donner l'ordre de grandeur. Toutefois, cette manière de procéder néglige trois facteurs importants : 1) les variations dans le temps des effectifs des PCS, alors que l'on sait par exemple que le nombre des agriculteurs continue à décroître, tandis que celui des professions intermédiaires augmente ; 2) les différences dans la durée des cursus, comme nous l'avons déjà signalé ; 3) les différences du nombre moyen d'enfants par PCS (fécondité différentielle). Si l'on désirait obtenir une mesure rigoureuse de ces différences, il faudrait

[1] À partir du tableau 2 de la page 159 de la même publication, on peut calculer les proportions de choix par filière. On obtient : 11,9 % pour le droit ; 9,8 % pour l'économie ; 34,8 % pour les "lettres" ; 20,7 % pour les sciences ; 7,2 % pour les disciplines de santé ; 15,6 % pour les IUT.

rapporter, pour chaque PCS, le nombre d'étudiants d'un âge donné (par exemple 20 ans) à l'effectif de l'ensemble de la classe d'âge.

La raison d'être de ce tableau est l'examen des relations statistiques entre la filière d'études et la profession des parents. Il serait possible de considérer les cases du tableau (marges exclues) comme une distribution unidimensionnelle, et de rechercher par exemple quelles sont les configurations les plus rares (il n'y a que 0,2 % des étudiants qui soient enfants d'agriculteurs et étudiants en médecine) ou les plus fréquentes (9,7 % sont étudiants en lettres issus d'une famille de cadre supérieur ou de profession libérale). Cette démarche, souvent fastidieuse, peut se justifier dans certains cas. Pour décrire et analyser les relations entre l'origine sociale et la filière d'études, plusieurs approches sont envisageables ; nous en examinerons trois : la comparaison avec une distribution "théorique", le calcul des pourcentages de ligne, et le calcul des pourcentages de colonne.

b) Comparaison avec un modèle théorique.

Une première démarche, qui est à la base de techniques statistiques classiques (comme le χ^2), consiste à confronter les données d'observation à des données construites dans une perspective théorique. Le modèle théorique le plus simple, et par conséquent le plus facile à construire, consiste à supposer qu'aucune relation n'existe entre les variables : dans ce cas précis, cela revient à faire l'hypothèse qu'il n'y a pas de relation entre la discipline de l'étudiant et son origine sociale. Si l'hypothèse de l'indépendance des variables était vraie, on observerait par exemple proportionnellement autant d'étudiants en droit chez les enfants d'agriculteurs que chez les enfants d'employés, d'ouvriers, ou de cadres supérieur ; c'est à dire proportionnellement autant d'étudiants en droit que dans la distribution marginale correspondant à l'ensemble de la population (totaux de colonnes). Par conséquent, si, sur les 1 302 177 étudiants du tableau 10, on compte 167 994 étudiants en droit, sur les 30 885 étudiants enfants d'agriculteurs, on en comptera : (167 994 / 1 302 177) × 30 885 ≅ 3 985. En appliquant ce raisonnement à chaque case du tableau 10, on obtient le tableau 11.

Origine socioprofessionnelle (PCS)	Droit	Économie	Lettres	Sciences et STAPS	Santé	IUT	Ensemble Université
Agriculteurs	3 985	3 171	10 849	7 313	2 988	2 579	30 885
Artisans, comm., chefs d'entreprise	12 784	10 173	34 809	23 463	9 588	8 276	99 093
Prof. libérales, cadres supérieurs	54 393	43 282	148 104	99 834	40 796	35 211	421 620
Professions intermédiaires	28 417	22 612	77 376	52 157	21 313	18 396	220 271
Employés	19 953	15 877	54 328	36 622	14 965	12 916	154 661
Ouvriers	18 822	14 977	51 249	34 545	14 117	12 184	145 894
Retraités, inactifs	13 435	10 691	36 581	24 659	10 076	8 697	104 139
Indéterminé	16 205	12 895	44 125	29 744	12 154	10 491	125 614
Ensemble	167 994	133 678	457 421	308 337	125 997	108 750	1 302 177

Tableau 11. Distribution théorique correspondant à l'hypothèse d'indépendance des variables.

Il est alors facile, en comparant les deux tableaux, d'évaluer dans quelle mesure les données observées s'écartent du modèle correspondant à l'hypothèse choisie. On peut par exemple calculer les "excédents" et les "déficits" en soustrayant le tableau 11 du tableau 10. On observe ainsi (tableau 12) que les plus forts "excédents" se trouvent dans la filière santé (enfants de cadres supérieurs et de PCS "indéterminée"), et les filières sciences (enfants de cadres supérieurs), droit (enfants de cadres supérieurs), et IUT (enfants d'ouvriers). Parallèlement, les plus forts déficits s'observent dans la filière lettres (enfants de cadres supérieurs), puis dans la filière santé (enfants d'ouvriers et d'employés) et les IUT (enfants de cadres supérieurs). Cette approche peut servir par exemple à donner une idée des "correctifs" qui seraient nécessaires pour une répartition plus "équitable" des flux d'étudiants selon les filières.

Origine socioprofessionnelle (PCS)	Droit	Économie	Lettres	Sciences et STAPS	Santé	IUT	Total
Agriculteurs	- 891	+ 166	- 935	+ 777	- 572	+ 1 455	0
Artisans, comm., chefs d'entreprise	+ 2 577	+ 1 945	- 2 016	- 2 081	- 1 619	+ 1 194	0
Prof. libérales, cadres supérieurs	+ 7 397	- 2 178	- 21 592	+ 9 703	+ 14 205	- 7 535	0
Professions intermédiaires	- 4 519	- 2 140	+ 4 072	+ 4 660	- 5 590	+ 3 517	0
Employés	+ 546	+ 1 114	+ 5 341	- 2 306	- 7 506	+ 2 811	0
Ouvriers	- 3 374	+ 1 871	+ 4 762	- 1 789	- 7 658	+ 6 188	0
Retraités, inactifs	+ 1 037	+ 1 885	+ 7 188	- 4 465	- 3 893	- 1 752	0
Indéterminé	- 2 773	- 2 663	+ 3 180	- 4 499	+ 12 633	- 5 878	0
Total	0	0	0	0	0	0	0

Tableau 12. Écarts des données observées par rapport à l'hypothèse d'indépendance des variables.

Le modèle correspondant à l'indépendance des variables n'est évidemment pas le seul possible. Il serait assez facile de construire d'autres tableaux "théoriques" concrétisant un raisonnement sur les mécanismes de répartition des étudiants dans les filières. Les seules contraintes sont : 1) la conservation des distributions marginales du tableau d'origine (afin que les données observées et les données construites restent comparables), et donc des totalisations en lignes et en colonnes ; 2) l'interdiction des effectifs négatifs (qui n'auraient ici aucun sens). On trouve de telles constructions dans les travaux sur la mobilité sociale par exemple (modèles de mobilité "en escalier" [1]).

L'intérêt de ces modèles est de faire le lien entre les données d'observation, et les théories sociologiques élaborées pour en rendre compte. En particulier, lorsque l'on doit choisir entre plusieurs explications possibles, il est utile de mesurer l'écart global entre les tableaux construits d'après ces théories, et le tableau d'origine ; plus l'écart sera faible, mieux

[1] Voir par exemple : Daniel Bertaux, "Sur l'analyse des tables de mobilité sociale", *Revue française de sociologie*, X (1969), 448-490.

la théorie sera adaptée aux observations. Pour évaluer l'écart global entre deux tableaux ayant les mêmes marges, on pourrait prendre, non la somme des écarts case par case (qui est nulle par construction), mais la somme des valeurs absolues de ces écarts. À cette mesure simpliste, on préfère un indice plus élaboré, mais conçu selon les mêmes principes, le χ^2 ("khi deux"). Cet indice tient compte de la valeur *relative* de l'écart (un petit écart sur un petit effectif est plus "grave" que le même écart sur un effectif important) ; en outre, il accentue les divergences entre les cases, en "pénalisant" les écarts les plus forts (élévation des écarts au carré).

c) Pourcentages de lignes : l'influence de l'origine socioprofessionnelle des étudiants.

L'importance des écarts entre le tableau 10 et le tableau 11 atteste l'existence d'une relation statistique entre l'origine sociale des étudiants et leur présence dans une filière d'études. On peut donc légitimement faire ici l'hypothèse d'une relation *causale*. Cette relation ne peut être que dans le sens d'une influence de l'origine socioprofessionnelle sur la présence dans une filière (et non l'inverse) ; mais l'on ne réfléchit généralement pas assez sur les fondements épistémologiques d'une telle affirmation.

En effet, les arguments en faveur de cette influence sont nombreux : l'origine sociale détermine dans une large mesure les moyens financiers de l'étudiant, et permet d'expliquer le choix (ou l'évitement) d'études longues, coûteuses, ou débouchant sur une profession qui nécessite une mise de fonds au départ ; la valorisation des diverses filières d'études n'est pas la même dans tous les milieux sociaux ; les acquis culturels indispensables pour certaines filières varient fortement d'un milieu à l'autre ; etc. Raisonner ainsi est faire ce que les logiciens appellent une *pétition de principe*, c'est à dire fonder son raisonnement sur la chose même que l'on prétend démontrer. Ce sont des explications intéressantes *si et seulement si* cette influence existe réellement.

Dans le cas le plus simple, le modèle de relation causale repose sur deux principes : 1) les effets ne peuvent précéder les causes ; 2) les variations des causes entraînent des variations dans les effets (ou encore : la disparition des causes entraîne la disparition des effets). Ces deux principes sont ici vérifiés : 1) on sait qu'au départ, les ressources matérielles des étudiants dépendent fortement des ressources des parents, même s'il existe des systèmes d'aides (bourses, allocations, prêts) destinés à atténuer ensuite les disparités dues au milieu d'origine ; et que les acquis culturels et les idéaux des étudiants jouent un rôle dans leur orientation même si certains déficits culturels peuvent ensuite être compensés. 2) le tableau 12 montre éloquemment qu'il existe des variations sensibles d'une filière à l'autre selon l'origine sociale des étudiants.

Pour analyser cette influence à partir du tableau d'origine (à deux dimensions), une méthode simple consiste à découper celui-ci en distributions à une seule dimension, que l'on comparera ensuite entre elles. La "variable explicative" (la cause supposée) étant l'origine socioprofessionnelle des étudiants, analyser son influence sur les filières d'études revient à comparer entre elles les lignes du tableau 10. Comme on ne peut comparer directement des sous-populations de tailles différentes (30 885 enfant d'agriculteurs, contre 421 620 enfants de cadres supérieurs), on calculera les pourcentages correspondant à chaque catégorie socio-professionnelle (sur le total de lignes du tableau 10). On obtient ainsi le tableau 13.

Origine socioprofessionnelle (PCS)	Droit	Économie	Lettres	Sciences et STAPS	Santé	IUT	Base : 100 % =
Agriculteurs	10,0 %	10,8 %	32,1 %	26,2 %	7,8 %	13,1 %	30 885
Artisans, comm., chefs d'entreprise	15,5 %	12,2 %	33,1 %	21,6 %	8,0 %	9,6 %	99 093
Prof. libérales, cadres supérieurs	14,7 %	9,7 %	30,0 %	26,0 %	13,0 %	6,6 %	421 620
Professions intermédiaires	10,8 %	9,3 %	37,0 %	25,8 %	7,1 %	10,0 %	220 271
Employés	13,3 %	11,0 %	38,6 %	22,2 %	4,8 %	10,2 %	154 661
Ouvriers	10,6 %	11,6 %	38,4 %	22,5 %	4,4 %	12,6 %	145 894
Retraités, inactifs	13,9 %	12,1 %	42,0 %	19,4 %	5,9 %	6,7 %	104 139
Indéterminé	10,7 %	8,1 %	37,7 %	20,1 %	19,7 %	3,7 %	125 614
Ensemble	12,9 %	10,3 %	35,1 %	23,7 %	9,7 %	8,4 %	1 302 177

Tableau 13. Influence de l'origine socioprofessionnelle sur les filières d'études supérieures.

Un premier survol du tableau 13 montre la diversité des "profils" de chaque filière. Il confirme le taux prépondérant d'étudiants "littéraires" dans toutes les strates définies par l'origine sociale, avec toutefois une sur-représentation des enfants de retraités ou d'inactifs (42,0 %), d'employés (38,6 %), et d'ouvriers (38,4 %), et une sous-représentation des enfants de professions libérales ou de cadres supérieurs (30,0 %). Dans la filière "sciences et STAPS", ce sont les enfants d'agriculteurs (26,2 %), de cadres supérieurs (26,0 %) et de professions intermédiaires (25,8 %) qui sont sur-représentés, alors que le déficit relatif touche surtout les enfants de retraités et inactifs (19,4 %). La filière "droit" est très contrastée, puisque les strates d'origine forment deux groupes relativement homogènes et bien distincts : quatre strates se situent au dessus de 13,0 % (artisans et assimilés, cadres supérieurs, retraités, employés) et les quatre autres en dessous de 11,0 % (agriculteurs, ouvriers, "indéterminés", professions intermédiaires). En revanche, la filière "économie" montre une distribution peu contrastée. Les "disciplines de santé" ont une distribution caractéristique ("courbe en J") si l'on ordonne les PCS en taux croissants : les taux progressent lentement, de 4,4 % (enfants d'ouvriers) à 8,0 % (enfants d'artisans), pour passer brusquement à 13,0 % (enfants de cadres supérieurs) puis à 19,7 % ("indéterminés"). Enfin, les IUT ont une composition bien diversifiée, avec une sur-représentation des enfants d'agriculteurs (13,1 %) et d'ouvriers (12,6 %).

En toute rigueur, pour lire ce tableau, il faudrait comparer chaque ligne à toutes les autres, ce qui imposerait 21 comparaisons de lignes, case par case. C'est pourquoi la lecture du tableau 13 peut être grandement facilitée par sa traduction graphique. Deux représentations sont possibles, selon que l'on met en abscisses les filières suivies, ou les PCS. La seconde représentation illustre la description que nous venons de faire : en portant les PCS en abscisses (schéma 4), on met en évidence par exemple le caractère "complémentaire" des proportions d'étudiants en lettres et en sciences (les courbes correspondantes varient en sens inverse), ou encore le peu de différences dans les proportions d'étudiants en sciences économiques selon leur origine sociale.

Naturellement, rien n'interdit (puisqu'il s'agit ici de deux variables nominales) de modifier l'ordre des catégories, de façon à mettre en évidence tel ou tel phénomène. On peut ainsi par exemple refaire le schéma 4 en ordonnant les PCS selon les pourcentages décroissants de filières longues ou coûteuses, des disciplines de santé aux IUT. Le choix de la présentation ne modifie pas les données ; c'est un simple problème de communication, qui dépend des intentions de l'auteur du commentaire, et du déroulement de son raisonnement.

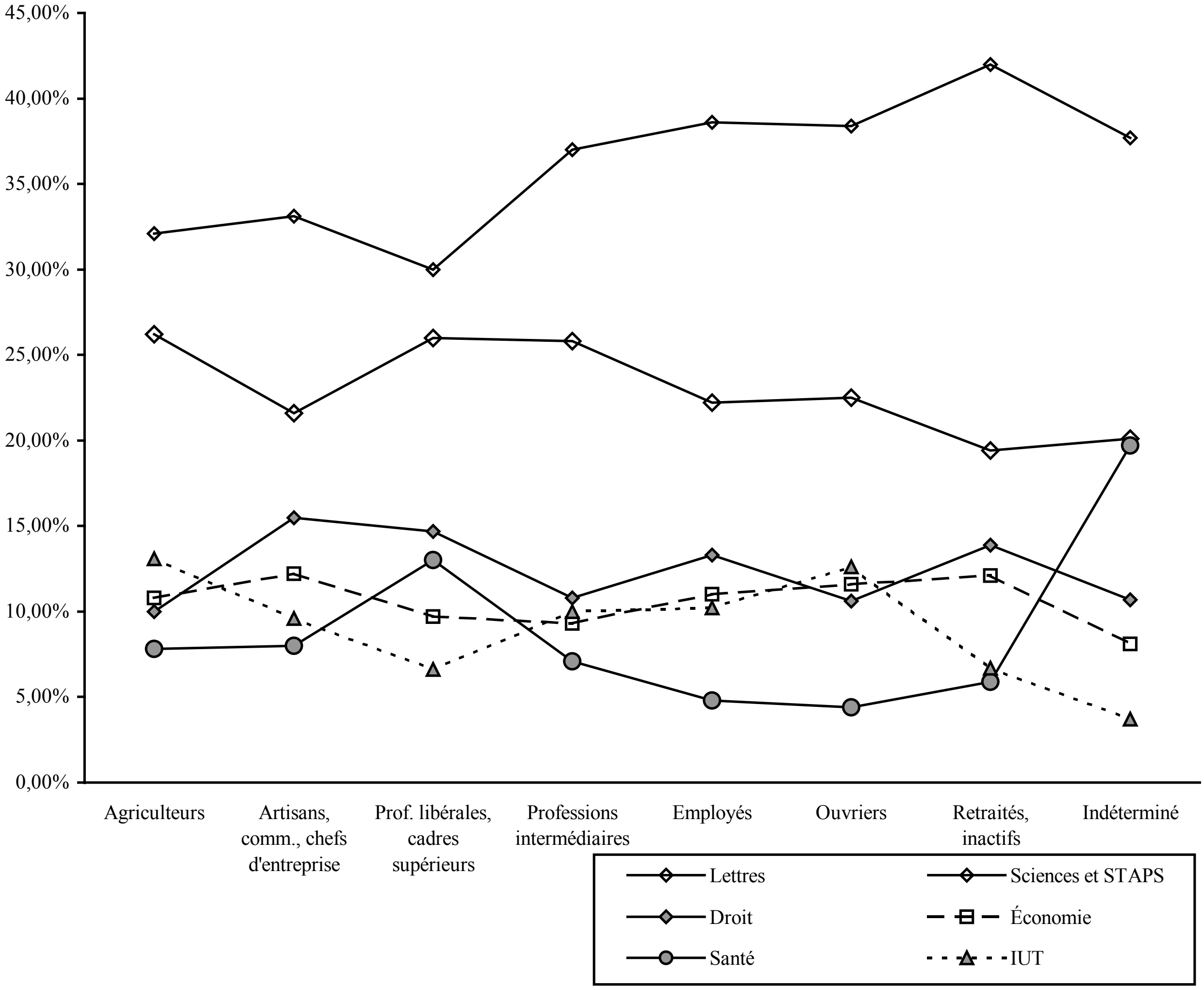


Schéma 4. Influence de l'origine socioprofessionnelle sur la filière d'études.

Une autre manière de lire le tableau 13 consiste à comparer chaque ligne à la ligne des totaux de colonnes. Cette distribution marginale représente en effet la distribution "moyenne" des pourcentages, toutes origines socioprofessionnelles confondues. Cela réduit à 8 les comparaisons de lignes : pour chacune d'elles, on recherchera les cases pour lesquelles la proportion observée est soit nettement supérieure, soit nettement inférieure à la proportion moyenne (marginale). Un moyen commode consiste (comme précédemment) à soustraire de chaque ligne du tableau les pourcentages marginaux. On obtient ainsi le tableau 14, qui décrit ces écarts.

Origine socioprofessionnelle (PCS)	Droit	Économie	Lettres	Sciences et STAPS	Santé	IUT
Agriculteurs	- 2,9 %	+ 0,5 %	- 3,0 %	+ 2,5 %	- 1,9 %	+ 4,7 %
Artisans, comm., chefs d'entreprise	+ 2,6 %	+ 1,9 %	- 2,0 %	- 2,1 %	- 1,7 %	+ 1,2 %
Prof. libérales, cadres supérieurs	+ 1,8 %	- 0,6 %	- 5,1 %	+ 2,3 %	+ 3,3 %	- 1,8 %
Professions intermédiaires	- 2,1 %	- 1,0 %	+ 1,9 %	+ 2,1 %	- 2,6 %	+ 1,6 %
Employés	+ 0,4 %	+ 0,7 %	+ 3,5 %	- 1,5 %	- 4,9 %	+ 1,8 %
Ouvriers	- 2,3 %	+ 1,3 %	+ 3,3 %	- 1,2 %	- 5,3 %	+ 4,2 %
Retraités, inactifs	+ 1,0 %	+ 1,8 %	+ 6,9 %	- 4,3 %	- 3,8 %	- 1,7 %
Indéterminé	- 2,2 %	- 2,2 %	+ 2,6 %	- 3,6 %	+ 10,0 %	- 4,7 %

Tableau 14. Écarts à la distribution marginale du tableau 13.

On notera que, dans le tableau 14, la somme des écarts de chaque ligne est nulle (aux effets d'arrondi près) ; c'est une propriété des écarts que nous avons déjà observée à propos du tableau 12. Par contre, cette propriété ne se vérifie pas à propos des sommes de colonnes : les *contraintes de marge* ne s'exercent ici que dans chaque ligne (nous reviendrons sur ce point au § 2.3.d). La différence essentielle entre ces deux tableaux est que, dans le tableau 12, il s'agit d'*écarts absolus* (en effectifs), tandis que le tableau 14 présente des *écarts relatifs* (pour lesquels les effets liés aux différences d'effectifs des catégories sociales ont été supprimés). On notera également que le sens des écarts est le même dans les deux tableaux : cela s'explique par le fait que, dans les deux tableaux, les écarts sont en réalité des écarts à l'hypothèse d'indépendance des variables. Cela est évident pour le tableau 12, qui a été construit pour cela. Si l'on calcule les pourcentages de ligne à partir du tableau 11 (hypothèse d'indépendance), on constate que chaque ligne affiche les mêmes pourcentages, et que ceux-ci sont égaux à ceux de la ligne de totalisation. En effet, si les variables n'ont aucune relation statistique (hypothèse d'indépendance), cela signifie que toutes les distributions relatives (lignes, mais aussi colonnes) sont identiques. En revanche, l'amplitude des écarts relatifs n'est pas la même que celle des écarts absolus : par exemple, dans le tableau 14, le plus grand écart positif s'observe pour les enfants de PCS "indéterminées" qui étudient les disciplines de santé (+ 10,0 %), et le plus grand écart négatif pour les enfants d'ouvriers étudiant ces mêmes disciplines (- 5,3 %). Ces deux tableaux ne font donc pas double emploi (ils ne décrivent pas exactement la même chose).

Les écarts relatifs du tableau 14 mettent en évidence les ressemblances et les contrastes entre les PCS. Ainsi, les enfants d'ouvriers et d'employés ont des "profils" assez voisins, même si leurs écarts les plus grands ne se situent pas dans les mêmes filières : leurs écarts à la distribution moyenne sont du même ordre pour quatre filières, les seules différences notables s'observant pour le droit et les IUT. Ce tableau montre également que deux PCS, les artisans et assimilés, et les professions intermédiaires, sont relativement proches de la distribution moyenne (la somme des valeurs absolues de leurs écarts relatifs est faible), tandis que les "indéterminés" s'en écartent fortement.

d) Pourcentages de colonnes : la composition de chaque filière selon l'origine socio-professionnelle des étudiants.

L'analyse de l'influence de l'origine socioprofessionnelle sur les filières dans lesquelles se trouvent les étudiants n'est pas la seule analyse possible. Si l'on désire décrire la composition, en termes d'origine sociale, des populations étudiantes selon la filière d'études, il faut revenir au tableau initial. En effet, pour déterminer dans quelle mesure les filières d'études sont "populaires" ou "élitistes", il nous faut aborder le tableau 10 sous un autre angle : la question posée n'est plus centrée sur les PCS, mais sur les filières. En appliquant le même raisonnement qu'au paragraphe précédent (§ 2.2.c), il nous faut maintenant découper le tableau 10 en colonnes, et calculer les pourcentages sur les totaux de colonnes (tableau 15). Si l'on compare rapidement les deux tableaux de pourcentages, le tableau 13 montrait que les enfants d'agriculteurs et d'ouvriers étaient proportionnellement plus nombreux en IUT que les enfants de cadres supérieurs ; le tableau 15 nous apprend que pourtant l'origine socioprofessionnelle majoritaire chez les étudiants d'IUT est la catégorie "professions libérales, cadres supérieurs". Cela est dû à l'importance de cette catégorie dans l'ensemble de la population des étudiants d'université. Les filières les plus "populaires" le sont donc fort peu.

Origine socioprofessionnelle (PCS)	Droit	Économie	Lettres	Sciences et STAPS	Santé	IUT	Ensemble Université
Agriculteurs	1,8 %	2,5 %	2,2 %	2,6 %	1,9 %	3,7 %	2,4 %
Artisans, comm., chefs d'entreprise	9,1 %	9,1 %	7,2 %	6,9 %	6,3 %	8,7 %	7,6 %
Prof. libérales, cadres supérieurs	36,8 %	30,7 %	27,7 %	35,5 %	43,7 %	25,4 %	32,4 %
Professions intermédiaires	14,2 %	15,3 %	17,8 %	18,4 %	12,5 %	20,1 %	16,9 %
Employés	12,2 %	12,7 %	13,0 %	11,1 %	5,9 %	14,5 %	11,9 %
Ouvriers	9,2 %	12,6 %	12,2 %	10,6 %	5,1 %	16,9 %	11,2 %
Retraités, inactifs	8,6 %	9,4 %	9,6 %	6,5 %	4,9 %	6,4 %	8,0 %
Indéterminé	8,0 %	7,7 %	10,3 %	8,2 %	19,7 %	4,2 %	9,6 %
Base : 100,0 % =	167 994	133 678	457 421	308 337	125 997	108 750	1 302 177

Tableau 15. Origine socioprofessionnelle des étudiants selon la filière d'études.

2.3. Récapitulation et compléments.

a) Comment lire un tableau à double entrée.

Un tableau à deux dimensions peut toujours être considéré comme une juxtaposition de tableaux à une seule dimension ; les principes de base pour les lire sont donc les mêmes. Lire un tableau, c'est exprimer en langage naturel toutes les informations chiffrées qu'il contient. En pratique, l'objectif est moins ambitieux : toute lecture consiste seulement à *extraire* du tableau une partie de ces informations, en fonction de l'intérêt qu'on lui attribue. Cela implique évidemment que l'on se pose des questions à son sujet.

Pour cela, il faut commencer par avoir une idée de l'objet du tableau. En règle générale, on commencera par vérifier avec le maximum de précision (à partir des commentaires qui l'accompagnent) sur quelle population il porte, et s'il s'agit de l'ensemble de la population visée ou seulement d'un échantillon (représentatif, ou non ?). On examinera ensuite la manière dont les données de base ont été recueillies et traitées pour constituer le tableau : définition des catégories, texte des questions posées (s'il s'agit d'une enquête par questionnaire), regroupements éventuels, calculs de coefficients, etc. Parvenir à une compréhension correcte des indications fournies par l'auteur, même si celles-ci sont rigoureuses et complètes, impose toujours une lecture attentive. À ce stade, il est important de comprendre quelles étaient les intentions de l'auteur lorsqu'il a présenté ce tableau (dans le cas général : recherche d'éléments permettant de supposer l'existence ou l'absence d'une relation causale entre les deux variables).

L'étape suivante est la description des distributions marginales (lorsque l'on dispose des effectifs, ou lorsque la base des pourcentages est indiquée). Il est important de bien cerner les caractéristiques de ces distributions pour plusieurs raisons. Tout d'abord, elles apportent une première information sur la population étudiée : ainsi, il aurait été intéressant de connaître la distribution par âge des célibataires étudiés par Durkheim, ne serait-ce que pour estimer la mortalité non liée au suicide. Ensuite, ces distributions, traduites en pourcentages, constituent un étalon (une "moyenne") pour décrire les distributions partielles. Enfin, la forme de ces distributions donne une indication sur l'ampleur possible de la liaison entre les deux variables (comme nous le verrons au § 2.3 d).

La troisième étape est la description des relations entre les deux variables, c'est à dire du contenu même du tableau. La méthode adoptée dépend directement des questions que l'on se pose au sujet de ce tableau : en quoi diffère la composition sociale des diverses filières d'études supérieures ? L'âge a-t-il une influence sur le taux de suicides ? Etc. À ce stade, l'éventail des questions est assez large, et les outils simples dont le sociologue dispose sont essentiellement les pourcentages (et leurs équivalents : fréquences relatives et "pour mille"), et les mesures de covariations (coefficients de corrélation, de contingence, d'association, etc.).

b) Sur quelle(s) base(s) calculer les pourcentages ?.

Les pourcentages sont un outil de comparaison très simple, et en général assez bien compris même par les non initiés (qui cependant ne parviennent pas toujours à éviter les erreurs d'interprétations). Dans un tableau à deux dimensions, deux principales options sont envisageables pour le choix de la base des pourcentages : les totaux de ligne, et les totaux de colonnes. Le choix entre ces deux options est simple, chaque fois que l'on peut faire l'hypothèse d'une relation causale entre les deux variables. Une telle relation n'a pas à être certaine *a priori* (c'est l'analyse du tableau qui établira son existence ou son absence), mais simplement possible (la cause supposée doit seulement précéder l'effet).

Si l'on admet la possibilité d'une telle relation, avec une "variable explicative", ou *cause* (*C*), et une "variable à expliquer", ou *effet* (*E*), l'interrogation la plus fréquente à propos du tableau est une question de la forme : y a-t-il une influence de *C* sur *E* ? Si oui, est elle forte ou faible ? Éventuellement : quelle est sa forme (est-elle la même partout) ? Pour vérifier l'existence d'une relation statistique entre *C* et *E*, et décrire cette relation, le moyen le plus simple est de découper le tableau en autant de distributions à une dimension qu'il y a de valeurs de *C*, comme dans l'exemple du tableau 13. Dans cet exemple, *C* (l'origine socioprofessionnelle) découpe le tableau en lignes : les bases des pourcentages seront par conséquent les totaux de ligne. Il suffira ensuite de comparer entre elles ces distributions unidimensionnelles (ou leur écart à la marge) : s'il existe entre elles des différences sensibles, il existe une relation statistique ; sinon, les variables sont dites "indépendantes". Si *C* découpe le tableau en colonnes, les bases des pourcentages seront naturellement les totaux de colonnes.

Une autre forme d'interrogation dans le cadre d'une hypothétique relation causale est une question "descriptive", du type de celle qui sous-tend le tableau 15. Dans ce cas, on compare entre elles des sous-populations homogènes selon *E* : quelle est l'origine sociale des étudiants en lettres (comparée à celle des étudiants d'autres disciplines) ? ou encore : quel est l'âge moyen des célibataires suicidaires (comparé à l'âge moyen des non suicidaires) ? Si *E* découpe le tableau en lignes, les bases des pourcentages seront les totaux de ligne ; si *E* découpe le tableau en colonnes (comme dans le tableau 15), les bases seront les totaux de colonnes.

Il peut arriver que les deux variables du tableau aient le même "statut épistémologique", c'est à dire qu'aucune ne peut être supposée être la cause de l'autre. C'est le cas par exemple lorsque l'on décrit une population en croisant l'âge et le sexe, ou en croisant les réponses à deux questions d'opinion. Dans une telle situation, selon ce que l'on désire mettre en évidence dans le tableau, on choisira comme base(s) le total général ("il y a *n* % de filles mineures dans la population", "*n* % des répondants se déclarent d'accord avec les deux propositions"), ou bien les totaux de ligne ou de colonne ("*x* % des filles et *y* % des garçons sont mineurs", "*x* % de ceux qui approuvent la proposition A désapprouvent la proposition B").

Enfin, rien ne s'oppose à ce que l'on choisisse comme base(s) des pourcentages des totaux partiels (correspondant à des regroupements), ou encore l'effectif d'une autre population (comme par exemple la population totale dans l'exemple du tableau 2), ou même une valeur "théorique" correspondant à une hypothèse particulière. La seule condition est que ce choix soit pertinent par rapport à ce que le sociologue veut mettre en évidence.

c) La mesure de la liaison entre deux variables.

On considère qu'il y a une relation statistique (liaison, corrélation) entre deux variables lorsque, si l'une varie, l'autre varie également ; c'est ce que l'on appelle les *covariations*. Plus la distribution (bidimensionnelle) observée est éloignée de l'hypothèse d'indépendance, plus la liaison entre les deux variables est forte. Il existe de nombreux coefficients pour mesurer la force de cette relation. Comme tout indice statistique, ses conditions d'utilisation dépendent des hypothèses faites sur les variables (niveaux de mesure).

Tous les manuels de statistique descriptive indiquent ces coefficients. On mentionnera au niveau nominal, les coefficients d'association ou de contingence (dont le coefficient φ, dérivé du χ^2), et la mesure de l'information transmise ; au niveau ordinal, les coefficients de corrélation par rang (τ de Kendall, ρ de Spearman) avec ex æquo ; au niveau métrique, le r de Bravais Pearson. Certains indices sont conçus pour des situations hybrides ; comme par exemple le cas où une variable est mesurable et l'autre seulement nominale (coefficient de régression c). La plupart de ces coefficients ont été "normalisés" : ils varient dans des limites connues, de 0 à 1 pour les coefficients d'association (de contingence), ou de - 1 à + 1 pour les coefficients de corrélation (les valeurs négatives traduisant une liaison *inverse* : lorsque l'une des variables croît, l'autre décroît). La valeur 0 correspond toujours à l'absence de liaison (indépendance des variables), et la valeur 1 à la liaison la plus forte possible ; la valeur - 1 correspond à la liaison inverse la plus forte possible.

Il est prudent de se référer aux ouvrages spécialisés ou de prendre l'avis d'un statisticien avant d'utiliser ces divers indices. En particulier, il est important de connaître les hypothèses sur lesquels ils sont construits [1], afin de décider s'ils conviennent aux hypothèses faites sur les variables, ainsi qu'aux objectifs du chercheur. D'autre part, il est très important de vérifier sur les données elles mêmes la signification des valeurs de ces coefficients, avant d'en tirer des conclusions. On peut constater, par exemple, qu'à un coefficient de corrélation élevé (tel que $r = 0{,}75$) correspond un nuage de points relativement peu concentré, traduisant une liaison plutôt moyenne.

d) Contraintes de marges et degrés de liberté.

Les conclusions que le sociologue tire de l'analyse des données se fondent généralement sur des comparaisons : comparaison entre les cases d'un tableau "empirique" (données d'observation, comme celles du tableau 10) et celles d'un tableau "théorique" (construit à partir d'une hypothèse, comme le tableau 11) ; ou entre les lignes ou les colonnes à l'intérieur d'un même tableau (exemples : tableaux 13 et 15). Dans ces deux situations, il se trouve confronté aux contraintes de marge.

[1] À titre d'exemple, le coefficient de corrélation le plus utilisé, le *r* de Bravais-Pearson, impose des conditions très précises quant aux propriétés des variables (mesurables), de leurs distributions (unimodales, variances partielles homogènes), de la fonction exprimant la liaison (linéaire), et de l'ajustement de la fonction aux points (symétrique, moindres carrés). Le sociologue peut légitimement estimer, après avoir pris l'avis du statisticien, que ces conditions ne pas adaptées à ses données ou à ses hypothèses.

Lorsque l'on construit un tableau "théorique", c'est à dire un tableau qui décrit "ce qui se passerait si les hypothèses étaient vérifiées", on raisonne toujours "à marges constantes". Pour que le tableau "empirique" puisse être confronté au tableau "théorique", il faut que les seules différences entre ces deux tableaux concernent l'hypothèse que l'on désire "vérifier" (ou "infirmer"). Ce sont donc sur les effectifs des cases du tableau, et sur eux seuls, que doit porter la comparaison : le modèle théorique est confronté aux données d'observation "toutes choses égales par ailleurs". La construction du tableau "théorique" est soumise à une contrainte de bon sens : la somme des chiffres "théoriques" (quelle que soit la manière dont ils sont été calculés, et même s'ils ont des valeurs négatives [1]) doit respecter les totaux de lignes et de colonnes du tableau "empirique". C'est ce dernier point que l'on appelle les *contraintes de marges*.

Dans l'exemple du tableau matérialisant l'indépendance entre la filière d'études et l'origine socioprofessionnelle des étudiants (tableau 11), on doit en principe calculer 48 valeurs "théoriques" : 8 lignes × 6 colonnes = 48 cases. En pratique, lorsque l'on a déterminé la valeur des 7 premiers chiffres de la première colonne, il ne reste aucune marge de liberté pour le chiffre restant : puisque le total de colonne est fixé, la valeur du dernier chiffre est déterminée par celle des 7 premiers (on pourrait le calculer en soustrayant les 7 premiers du total de la colonne). On dira que cette colonne a 7 *degrés de liberté*. De même, lorsque l'on a fixé la valeur des chiffres des 5 premières colonnes, tous les chiffres de la sixième sont déterminés par les totaux de ligne. Ainsi, seulement 8 - 1 = 7 lignes et 6 - 1 = 5 colonnes sont affranchies des contraintes de marge : seules, 35 cases peuvent être déterminées librement. On dira que ce tableau a 35 degrés de liberté.

Plus généralement, dans les cas usuels, une distribution unidimensionnelle qui prend N valeurs a $N - 1$ degrés de liberté : lorsque l'on calcule des pourcentages sur le total, la valeur du dernier pourcentage est imposée (pour que le total fasse bien 100). Un tableau de L lignes et C colonnes a $(L - 1) \times (C - 1)$ degrés de libertés. Cette notion de *degrés de libertés* est très importante en statistique inférentielle : elle est nécessaire en particulier pour utiliser correctement certaines tables (comme celle du χ^2), dans lesquelles chaque ligne correspond au nombre de degrés de liberté de la distribution (symbolisé par la lettre grecque *nu* : ν).

Par ailleurs, les *contraintes de marge* ont une influence sur la forme et la force de la liaison entre les variables : plus la dispersion des distributions marginales est forte, plus la valeur maximum possible du coefficient mesurant la force de la liaison peut être élevée (et se rapprocher du "maximum théorique" fixé à + 1 ou - 1). Par contre, lorsque les effectifs des distributions marginales sont fortement concentrés sur quelques catégories seulement, le coefficient mesurant la liaison entre les variables a une valeur maximum possible inférieure à 1 (il ne varie pas dans l'intervalle [0 , 1], ou [- 1 , + 1], mais dans un intervalle plus petit) ; il est donc difficile dans ces conditions de comparer deux distributions différentes à partir de ce seul coefficient. C'est pourquoi il est prudent, avant de tirer des conclusions à partir des indices statistiques quels qu'ils soient, de toujours "retourner aux données", et contrôler directement sur les tableaux la *signification* qu'il convient de donner aux valeurs prises par ces indices.

[1] Dans les cas courants, on considère que tous les chiffres "théoriques" doivent être positifs ou nuls ; il existe pourtant des modèles dans lesquels un effectif négatif a un sens (il exprime un déficit, par exemple).

Chapitre 3. Distributions à plus de deux variables.

Les distributions à trois variables ou plus peuvent être considérées comme une généralisation des tableaux à deux dimensions, et leur lecture obéit en gros aux mêmes règles : il est toujours possible de les décomposer en tableaux plus simples, à une ou deux dimensions seulement. Cependant, il faut un minimum de pratique pour manipuler une distribution multidimensionnelle, ne serait-ce que pour effectuer les totalisations correspondant à chacune des variables considérée isolément. En outre, de même que les tableaux à deux dimensions nous ont conduit à présenter la notion de liaison statistique entre deux variables, les tableaux à plus de deux dimensions nous amènent à introduire celle, aussi fondamentale, d'*interaction* entre variables.

3.1. Premier exemple : influence de la classe d'origine et de la classe d'appartenance sur le sentiment d'appartenir à une classe sociale.

Le tableau 16 [1] présente les réponses des hommes chefs de ménage d'un échantillon national représentatif de 1 780 personnes, interrogées par l'IFOP en décembre 1966, sur le thème "l'univers politique des Français et l'image du parti Communiste". Les trois variables prises en compte dans ce tableau sont la profession de la personne interrogée, la profession de son père, et le sentiment d'appartenance à une classe sociale. Pour simplifier la présentation des résultats, la variable "profession" a été regroupée en trois catégories : ouvriers, agriculteurs (exploitants et salariés), et l'ensemble des autres professions. La variable "sentiment d'appartenance" a été construite à partir de deux questions enchaînées : "*Avez-vous le sentiment d'appartenir à une classe sociale ?*" ; si oui : "*Laquelle ?*" (question ouverte). Les réponses ont été regroupées en trois catégories : pas de sentiment d'appartenance ("non" à la première question), sentiment d'appartenir à la classe ouvrière (réponses à la question ouverte : "la classe ouvrière", "les ouvriers", ou encore "le prolétariat"), et sentiment d'appartenir à une autre classe (réponses : "les pauvres", "les travailleurs", "les Français moyens", "les fonctionnaires", etc.). Les personnes prises en compte dans ce tableau sont les hommes chefs de ménage ayant répondu aux deux questions, moins les agriculteurs.

L'objectif de ce tableau est de déceler une influence conjointe de la classe d'origine (dont l'indicateur est la profession du père) et de la classe d'appartenance (dont l'indicateur est la profession de la personne interrogée), sur le sentiment d'appartenance. Pour cela, on a ramené le tableau à un ensemble de distributions unidimensionnelles, et calculé les pourcentages relatifs au sentiment d'appartenance ("variable à expliquer") pour chacune des sous-populations définies par le croisement des deux autres variables ("variables explicatives").

De prime abord, on remarque que la proportion de ceux qui déclarent n'avoir pas de sentiment d'appartenance varie peu (de 28 % à 36 %) ; on note également que, pour ceux qui ont le sentiment d'appartenir à une classe sociale, l'influence de la classe d'appartenance sur la classe à laquelle on s'identifie est plus forte que l'influence de la classe d'origine. Mais, pour

[1] Tableau 51, page 220 de : Guy Michelat et Michel Simon, *Classe, religion et comportement politique*, Paris, Fondation Nationale des Sciences Politiques et Éditions Sociales, 1977.

affiner cette analyse, il serait utile de disposer des distributions marginales de chacune des trois variables du tableau, ainsi que des trois tableaux croisés correspondant à chaque combinaison deux à deux de ces variables. Pour cela, il est nécessaire de retrouver les effectifs à partir desquels les pourcentages ont été calculés.

Profession de la personne interrogée	Profession du père	Sentiment d'appartenance		Pas de sentiment d'appartenance	Base : 100 % =
		Classe ouvrière	Autres classes		
Ouvrier	Ouvrier	53 %	19 %	28 %	158
	Agriculteur	47 %	24 %	29 %	45
	Autres	47 %	23 %	30 %	64
Autres (sauf agriculteurs)	Ouvrier	28 %	39 %	32 %	117
	Agriculteur	16 %	48 %	36 %	73
	Autres	13 %	53 %	34 %	253

Tableau 16. Classe subjective en fonction de la profession de la personne interrogée et de la profession du père (hommes chefs de ménage).

a) Reconstitution du tableau d'effectifs ayant servi à calculer les pourcentages.

Pour constituer les distributions à une ou deux variables permettant une analyse plus approfondie du tableau, il suffit d'additionner judicieusement les effectifs du tableau à trois dimensions. Ainsi, pour obtenir la distribution de la profession du père, on additionnera les effectifs des fils d'ouvriers (158 + 117 = 275), d'agriculteurs (45 + 73 = 118), ou d'autres professions (64 + 253 = 317) ; on obtiendra ce faisant l'effectif des personnes figurant dans le tableau (275 + 118 + 317 = 710). D'autre part, le tableau croisé "profession de la personne interrogée par profession du père" est donné dans la colonne "Base = 100" du tableau d'origine ; il suffit pour le constituer de disposer côte à côte les deux moitiés de cette colonne, et d'effectuer les totalisations marginales.

En revanche, il n'est pas possible de constituer les tableaux "sentiment d'appartenance" croisé par "profession de la personne interrogée", et par "profession du père", sans avoir au préalable reconstitué les effectifs qui ont servi à calculer les pourcentages (Tableau 17). Par exemple, le nombre d'ouvriers fils d'ouvriers s'identifiant à la classe ouvrière est égal à 53 % de 158, soit : 158 × 53 / 100 = 83,74, que l'on arrondira à 84. En toute rigueur, compte tenu des calculs d'arrondi des pourcentages, il faudrait dire que cet effectif est compris entre 82,95 (soit 52,5 % de 158) et 84,53 (soit 53,49 % de 158), ce qui nous laisse le choix entre 83 et 84. En règle générale, plus la base des pourcentages est grande, moins l'estimation des effectifs correspondants est précise. C'est pourquoi il est conseillé de toujours commencer ces estimations par les effectifs correspondant aux bases les plus faibles, pour lesquels l'incertitude est moindre ; on ajustera ensuite, en fonction des contraintes de marges, les effectifs calculés sur les bases plus importantes. Ajoutons que, dans ce cas précis, les incertitudes dans les estimations n'ont pas d'effet sensible sur les analyses qui suivent : quels que soient les choix qui sont faits en matière d'arrondi, les conclusions de l'analyse restent les mêmes.

Profession de la personne interrogée	Profession du père	Sentiment d'appartenance		Pas de sentiment d'appartenance	Total
		Classe ouvrière	Autres classes		
Ouvrier	Ouvrier	84	30	44	158
	Agriculteur	21	11	13	45
	Autres	30	15	19	64
	Total	135	56	76	267
Autres (sauf agriculteurs)	Ouvrier	33	46	38	117
	Agriculteur	12	35	26	73
	Autres	33	134	86	253
	Total	78	215	150	443
Total toutes professions, sauf agriculteurs	Ouvrier	117	76	82	275
	Agriculteur	33	46	39	118
	Autres	63	149	105	317
	Total	213	271	226	710

Tableau 17. Estimation des effectifs ayant servi à constituer le tableau 16.

b) L'influence de la classe d'origine et de la classe objective sur le sentiment d'appartenance.

Les questions auxquelles ces données permettent de répondre peuvent être ainsi formulées : la classe d'origine et la classe objective influent-elles sur le sentiment d'appartenance, et si oui, laquelle influe le plus ? Le tableau 18 est constitué directement à partir des lignes "total" du tableau 17 (l'origine sociale n'est pas prise en considération). La première colonne du tableau 18 correspond à la première ligne "total" du tableau 17 : les 135 ouvriers qui ont le sentiment d'appartenir à la classe ouvrière représentent 51 % des 267 ouvriers, et ainsi de suite. On vérifiera que les deux autres colonnes correspondent aux deux autres lignes "total". Ce tableau met en évidence la forte influence de la position sociale sur le sentiment d'appartenance.

Sentiment d'appartenance	Classe d'appartenance		Ensemble
	Ouvrier	Autre profession	

Classe ouvrière Autre classe Pas de sentiment d'app.	51 % 21 % 28 %	18 % 48 % 34 %	30 % 38 % 32 %
Base : 100 % =	267	443	710

Tableau 18. Sentiment d'appartenance selon la classe d'appartenance (pourcentages).

Le tableau 19 correspond au dernier tiers du tableau 17 (total toutes professions confondues). Il révèle une influence également sensible de la classe d'origine sur le sentiment d'appartenance à la classe ouvrière.

Sentiment d'appartenance	Classe d'origine			Ensemble
	Père ouvrier	Père agriculteur	Père autre prof.	
Classe ouvrière	42 %	28 %	20 %	30 %
Autre classe	28 %	39 %	47 %	38 %
Pas de sent. app.	30 %	33 %	33 %	32 %
Base : 100 % =	275	118	317	710

Tableau 19. Sentiment d'appartenance selon la classe d'origine.

Il apparaît donc que le sentiment d'appartenir à la classe ouvrière est plus fort si l'on est soit fils d'ouvrier (Tableau 19), soit ouvrier soi-même (Tableau 18) ; mais à ce stade de l'analyse, nous ne pouvons pas dire avec certitude laquelle de ces deux influences est la plus forte, ni si les influences respectives se renforcent ou, au contraire, s'annulent. Cette dernière hypothèse est d'autant moins à écarter que la classe d'appartenance est manifestement liée à la classe d'origine (Tableau 20) ; il est donc possible qu'il s'agisse en réalité des deux faces d'une seule et même influence. Notons que le tableau 20 est construit à partir de la colonne "total" du tableau 17, la première ligne du tableau 20 correspondant au premier tiers de la colonne "total" du tableau 17 (ouvriers distingués selon leur classe d'origine), et ainsi de suite.

Classe d'appartenance	Classe d'origine			Ensemble
	Père ouvrier	Père agriculteur	Père autre prof.	
Ouvrier	57 %	38 %	20 %	38 %
Autre profession	43 %	62 %	80 %	62 %
Base : 100 % =	275	118	317	710

Tableau 20. Classe d'appartenance selon la classe d'origine.

c) Deux démarches pour l'analyse globale du tableau.

Le tableau 16 ne se réduit évidemment pas aux trois tableaux croisés ci-dessus. Comment interpréter ce tableau dans sa totalité ? L'obstacle à la lecture globale du tableau tient à sa relative complexité. Aussi, plutôt que de chercher à extraire de ce tableau toutes les informations possibles, est-il préférable de poser clairement une série de questions auxquelles ces données devraient permettre de répondre. En règle générale, plus les données sont complexes, plus il est nécessaire de les aborder sous l'angle de problèmes particuliers.

Une fois admis le schéma explicatif proposé par les auteurs, il est intéressant de rechercher quelle est l'influence simultanée (ou conjointe), sur le sentiment d'appartenance, de la classe d'origine et de la classe objective. Il y a deux manières simples de répondre à cette question : on peut combiner les deux "variables explicatives" et les considérer comme une seule variable causale, ou l'on peut spécifier la question en la précisant.

La première démarche est celle adoptée par les auteurs : chaque ligne du tableau 16 correspond à un état de la variable composite obtenue en combinant les deux variables supposées agir sur le sentiment d'appartenance. On peut, par lecture directe du tableau d'origine, faire deux principaux constats : 1) lorsque les traits d'appartenance objective à la classe ouvrière diminuent, le taux d'existence d'un sentiment d'appartenance à une classe quelconque décroît légèrement (de 100 % - 28 % = 72 % à 100 % - 36 % = 64 %) et assez régulièrement ; 2) parallèlement, le sentiment d'appartenir à la classe ouvrière décroît fortement (de 53 % à 13 %), et également régulièrement. On peut en inférer que la classe objective a, sur le sentiment d'appartenance à la classe ouvrière, un effet massif ; et que, toutes choses égales par ailleurs, la classe d'origine a un effet moindre, mais de même sens, qui s'ajoute par conséquent à l'effet principal.

Les auteurs ont d'ailleurs choisi la présentation du tableau 16 afin de mettre en évidence cette liaison statistique. En effet, s'ils avaient ordonné les états de la variable composite en commençant par la classe d'origine (profession du père), les taux correspondants d'appartenance à la classe ouvrière se seraient présentés dans l'ordre suivant :

53 %   28 %   47 %   16 %   47 %   13 %

Bien que les chiffres soient les mêmes, le sens de la relation n'est plus aussi apparent.

Les données statistiques ne présentent pas toujours des relations entre variables aussi claires. Aussi est-il plus commode, dans le cas général, d'"interroger les données", en posant à leur propos des questions spécifiques (c'est-à-dire en formulant des hypothèses). Par exemple, dans quelle mesure le fait d'être ouvrier, ou fils d'ouvrier, influe-t-il sur le sentiment d'appartenir à la classe ouvrière ? À partir du tableau 17, il est aisé de constituer le tableau permettant de répondre à cette question (pour mettre en évidence la manière dont le tableau 21 a été construit, nous avons conservé les effectifs correspondant aux pourcentages).

Profession de la personne interrogée	Profession du père	Sentiment d'appartenir à la classe ouvrière	Autres réponses	Total
Ouvrier	Ouvrier	84 = 53 %	74 = 47 %	158 = 100 %
	Autre	51 = 47 %	58 = 53 %	109 = 100 %
Autre (sauf agriculteur)	Ouvrier	33 = 28 %	84 = 72 %	117 = 100 %
	Autre	45 = 14 %	281 = 86 %	326 = 100 %
Ensemble (sauf agriculteurs)		213 = 30 %	497 = 70 %	710 = 100 %

Tableau 21. Sentiment d'appartenance à la classe ouvrière,
selon que l'on est ouvrier ou fils d'ouvrier.

Notons que, pour répondre à une question précise, on est le plus souvent amené à simplifier le tableau, en procédant à des regroupements. Afin de mettre en évidence la part de la classe d'appartenance et de la classe d'origine sur le sentiment d'appartenir à la classe ouvrière, nous avons dans ce tableau distingué quatre sous-populations, selon qu'elles étaient ou non soumises à chacune des causes étudiées. Pour lire le tableau 21, il est commode de commencer par la sous-population qui n'est soumise à aucune de ces causes : les 326 hommes qui ne sont ni ouvriers, ni fils d'ouvrier. Parmi ceux-ci, seulement 14 % s'identifient à la classe ouvrière ; ce taux est par conséquent dû à d'autres facteurs que ceux retenus, et cette sous-population va en quelque sorte servir de groupe témoin (ou, si l'on préfère, de "point zéro") pour évaluer la force des effets recherchés. L'influence de la classe d'origine seule porte ce taux à 28 %, soit un accroissement de 14 % ; celle de la classe d'appartenance seule le porte à 47 %, soit un accroissement de 33 %. L'influence de la classe d'appartenance est donc de loin la plus importante. Enfin, l'influence simultanée de ces deux facteurs porte le taux étudié à 53 %, soit un accroissement de 39 %. Cet accroissement est le plus élevé des trois, mais il est moindre que la somme des accroissements dus aux effets de chaque variable prise isolément. Il correspond à ce qu'on appelle un effet d'interaction.

d) Les effets d'interaction.

Nous avons vu l'importance de la notion de liaison statistique dès que deux variables sont mises en relation ; à partir de trois variables apparaît en outre la notion d'*interaction*. Cette notion (mise en évidence par l'analyse de la variance) correspond à l'influence conjointe, sur une "variable à expliquer", de deux "variables explicatives" ou plus. Pour simplifier notre présentation, nous allons supposer que les effets des "variables explicatives" s'ajoutent (hypothèse de l'additivité des facteurs, qui est à la base de l'analyse de la variance). Dans l'exemple du tableau 21, à l'effet de facteurs non présents dans les données (+ 14 %) pourraient s'ajouter les effets de la classe d'origine (+ 14 %) et de la classe d'appartenance (+ 33 %). Si ces effets pouvaient effectivement s'ajouter, le taux d'identification à la classe ouvrière pour les ouvriers fils d'ouvriers devrait être de : 14 % + 14 % + 33 % = 61 % ; or, le taux réellement observé n'est que de 53 %. Pour conserver l'hypothèse commode de l'additivité des effets, nous devons introduire un facteur de correction, appelé *effet* *d'interaction*, qui se trouve ici être négatif, et valoir - 8 % (14 % + 14 % + 33 % **- 8 %** = 53 %).

Profession de la personne interrogée	Profession du père	Sentiment d'appartenir à une classe sociale	Autres réponses	Total
Ouvrier	Ouvrier	114 = 72 %	44 = 28 %	158 = 100 %
	Autre	77 = 71 %	32 = 29 %	109 = 100 %
Autre (sauf agriculteur)	Ouvrier	79 = 68 %	38 = 32 %	117 = 100 %
	Autre	214 = 66 %	112 = 34 %	326 = 100 %
Ensemble (sauf agriculteurs)		484 = 68 %	226 = 32 %	710 = 100 %

Tableau 22. Sentiment d'appartenance à une classe sociale,
selon que l'on est ouvrier ou fils d'ouvrier.

Cet effet d'interaction peut d'ailleurs, selon les cas, être également positif ou nul. Par exemple, si nous désirons mesurer l'influence de ces mêmes "variables explicatives" sur le sentiment d'appartenance à une classe sociale quelle qu'elle soit, il est facile de constituer, à partir du tableau 17, le tableau permettant de répondre à cette question (tableau 22). La décomposition en effets additifs des taux observés donne un effet résiduel dû à d'autres facteurs très élevé (+ 66 %), des effets relativement faibles dus à la classe d'appartenance (+ 5 %) et à la classe d'origine (+ 2 %), et un effet d'interaction pratiquement nul (- 1 %).

D'autre part, dans l'analyse des réponses à cette enquête, les auteurs examinent l'influence de ces mêmes "variables explicatives" sur les intentions de vote [1]. En procédant pour ce tableau comme nous l'avons fait précédemment, on obtient, après élimination des non réponses, les taux suivants d'intention de vote pour un parti de gauche :

- ouvrier fils d'ouvrier : 68 %
- ouvrier d'autre origine sociale : 58 %
- non ouvrier fils d'ouvrier : 48 %
- non ouvrier d'autre origine sociale : 41 %

La décomposition de ces taux observés en effets additifs montre qu'aux effets résiduels (+ 41 %) s'ajoutent non seulement l'effet de la classe d'appartenance (+ 17 %) et celui de la classe d'origine (+ 7 %), mais aussi un effet d'interaction faible, mais positif (+ 3 %). Dans cet exemple, le fait d'être à la fois ouvrier et fils d'ouvrier majore les effets isolés de la classe d'appartenance et de la classe d'origine : il y a en quelque sorte un effet de synergie qui accroît encore la probabilité de voter à gauche pour les ouvriers fils d'ouvrier (interaction positive).

Dans le premier exemple par contre, l'interaction négative signifie que, s'il est vrai que l'on a plus le sentiment d'appartenir à la classe ouvrière lorsque l'on est à la fois ouvrier et fils

[1] *Op. cit.*, tableau 14, page 166.

d'ouvrier, il y a ici redondance entre les influences de la classe d'origine et de la classe d'appartenance sur ce sentiment identitaire. Le second exemple enfin illustre le fait qu'il peut ne pas y avoir d'effet d'interaction. Pour conclure, remarquons que le fait que les "variables explicatives" sont liées entre elles (tableau 18) ne détermine pas le sens de l'effet d'interaction, qui peut être positif, négatif, ou nul selon la "variable à expliquer".

3.2. Deuxième exemple : la réussite scolaire des étudiants de première année d'études littéraires à Paris en 1962-63.

Le tableau 23 est tiré d'une étude sur les facteurs de réussite des étudiants en première année de Lettres et Sciences Humaines à la Sorbonne, en 1963 [1]. C'est un tableau à six dimensions : cinq "variables explicatives", dont quatre dichotomiques (section suivie dans l'enseignement secondaire, mention au Bac, âge à l'entrée dans l'enseignement supérieur, exercice d'une profession parallèlement aux études) et une à trois états (profession du père, servant d'indicateur de l'origine sociale), plus une "variable à expliquer", la réussite aux sessions de juin et d'octobre 1963 du CELG (examen correspondant alors à la première année du DEUG). La présentation adoptée par l'auteur de l'article répond à un souci de mise en page ; d'autres présentations auraient naturellement été possibles. Pour chacune des 48 sous-populations définies par la combinaison des cinq "variables explicatives", seuls figurent les taux de réussite, suivis de l'effectif sur lequel ils ont été calculés. Ces effectifs sont très dissemblables, puisqu'ils vont de 2 à 670 selon les configurations de réponses ; aussi, afin de limiter les risques d'interprétation "sauvage" (dus à un survol trop hâtif du tableau), l'auteur a préféré ne pas présenter de pourcentages pour des effectifs inférieurs à 10.

[1] Page 196 de : Noëlle Bisseret, "La 'naissance' et le diplôme. Les processus de sélection au début des études universitaires", *Revue française de sociologie*, *IX* (1968), 185-207.

Exerce une profession	Âge à la première inscription	Mention au Bac	Études secondaires	Origine sociale: Agriculteurs, ouvriers, petits employés	Origine sociale: Artisans, commerçants, cadres moyens	Origine sociale: Industriels, prof. libérales, cadres sup.
non	< 20 ans	mention	classiques	86 % (42)	88 % (108)	89 % (297)
non	< 20 ans	mention	modernes	69 % (58)	79 % (122)	67 % (161)
non	< 20 ans	non	classiques	60 % (82)	70 % (166)	68 % (520)
non	< 20 ans	non	modernes	45 % (203)	54 % (384)	52 % (670)
non	≥ 20 ans	mention	classiques	3/3	6/8	68 % (31)
non	≥ 20 ans	mention	modernes	35 % (26)	43 % (44)	45 % (53)
non	≥ 20 ans	non	classiques	37 % (27)	47 % (45)	50 % (147)
non	≥ 20 ans	non	modernes	40 % (114)	45 % (181)	39 % (411)
oui	< 20 ans	mention	classiques	4/5	2/2	1/2
oui	< 20 ans	mention	modernes	51 % (35)	62 % (21)	4/8
oui	< 20 ans	non	classiques	3/7	27 % (15)	29 % (17)
oui	< 20 ans	non	modernes	23 % (86)	33 % (73)	30 % (36)
oui	≥ 20 ans	mention	classiques	64 % (11)	48 % (23)	11 % (18)
oui	≥ 20 ans	mention	modernes	36 % (123)	28 % (130)	31 % (105)
oui	≥ 20 ans	non	classiques	23 % (69)	26 % (64)	23 % (111)
oui	≥ 20 ans	non	modernes	19 % (387)	23 % (400)	24 % (397)

Tableau 23. Taux de réussite au Certificat d'Études Littéraires Générales
en 1963 à Paris, selon divers facteurs.

La population étudiée est constituée des 6 919 étudiants inscrits en première année à la Faculté des Lettres et Sciences Humaines de Paris (Sorbonne) en 1962-63. La réussite est mesurée à partir des résultats des deux sessions d'examens de 1963. Le taux de réussite global par rapport aux inscrits est de 46 % ; les 54 % restants se décomposent en abandons (34 %) et redoublements (20 %). L'effectif du tableau 23 est de 6 048 étudiants seulement, l'auteur ayant éliminé les individus pour lesquels manquait au moins une des informations requises. Sur cette sous-population, le taux de réussite global est de 47 %.

a) Comment lire le tableau 23 ?

Ce tableau dépasse largement en complexité le tableau 16. Si l'on désire l'analyser aussi complètement que possible, il est conseillé de se reporter aux données d'origine, c'est à dire aux effectifs, voire de les estimer si l'on n'en dispose pas (ce qui est le cas ici). À partir de ce tableau, il est aisé de reconstituer (comme pour le tableau 16) les dénombrements qui ont permis de le calculer (Tableau 24). Seul ce tableau d'effectifs nous permettra de répondre à l'ensemble des questions que l'on peut se poser à propos de ces données. En effet, en disposant des effectifs correspondant à chaque "patron de réponses" (*pattern* = configuration), nous pouvons constituer tous les tableaux à deux, trois, quatre ou cinq variables nécessaires pour étudier par exemple l'effet de l'origine sociale sur la section choisie dans l'enseignement secondaire ou sur l'obtention d'une mention au Bac, ou encore les effets conjugués sur la réussite au CELG de l'âge à l'entrée dans l'enseignement supérieur et de l'exercice d'une profession. Pour simplifier la lecture de ce tableau, nous avons regroupé les professions en deux catégories d'effectifs comparables, sur la base de leurs niveaux économique et culturel moyens.

Cependant, le tableau 23 a été conçu pour répondre à une question plus générale : des cinq causes supposées (figurant dans le tableau), laquelle influe le plus fortement sur la réussite à l'examen ? Quelle est celle dont l'influence se manifeste le plus ensuite ? Comment se traduisent les éventuelles interactions entre ces variables ? Comme le fait remarquer l'auteur, "d'une part, il se produit un effet cumulatif des variables, et d'autre part, il s'établit un jeu de compensations qui veut que l'influence de l'une ne peut se comprendre que par sa combinaison avec les autres" (*ibid.*). Une démarche particulièrement adaptée à l'analyse des influences directes et des interactions d'un ensemble de variables sur une "variable à expliquer" dichotomique est la procédure de *segmentation* ; nous allons en décrire le principe au paragraphe suivant.

b) Quels sont les facteurs de la réussite au CELG ?

Pour déterminer quelle est la variable dont la réussite à l'examen dépend le plus fortement, il suffit de constituer les cinq tableaux croisant chacune des "variables explicatives" avec la réussite au CELG. Pour comparer la force de la liaison entre chaque "variable explicative" et le taux de réussite, de nombreux coefficients sont disponibles (coefficient de contingence, information transmise, etc.) ; chacun a des propriétés différentes. Nous utilisons ici le plus simple à calculer, qui est la différence des pourcentages de réussite et d'échec. Il faut toutefois se souvenir qu'en règle générale, l'usage d'un coefficient différent peut conduire à des résultats différents, ce qui serait le cas ici.

Origine sociale	Travail salarié	Âge 1ère inscription	Mention au Bac	Section du secondaire	Réussite	Abandon, redoubl.	Total
Agriculteur, ouvrier, petit employé, artisan, commerçant, cadre moyen	non	moins de 20 ans	oui	classique	131	19	150
				moderne	136	44	180
			non	classique	165	83	248
				moderne	299	288	587
		20 ans et au-dessus	oui	classique	9	2	11
				moderne	28	42	70
			non	classique	31	41	72
				moderne	127	168	295
	oui	moins de 20 ans	oui	classique	6	1	7
				moderne	31	25	56
			non	classique	7	15	22
				moderne	44	115	159
		20 ans et au-dessus	oui	classique	18	16	34
				moderne	81	172	253
			non	classique	33	100	133
				moderne	166	621	787
Industriel, profession libérale, cadre supérieur	non	moins de 20 ans	oui	classique	264	33	297
				moderne	108	53	161
			non	classique	354	166	520
				moderne	348	322	670
		20 ans et au-dessus	oui	classique	21	10	31
				moderne	24	29	53
			non	classique	74	73	147
				moderne	160	251	411
	oui	moins de 20 ans	oui	classique	1	1	2
				moderne	4	4	8
			non	classique	5	12	17
				moderne	11	25	36
		20 ans et au-dessus	oui	classique	2	16	18
				moderne	33	72	105
			non	classique	26	85	111
				moderne	95	302	397
Ensemble des étudiants candidats au CELG en 1963					2 842	3 206	6 048

Tableau 24. Effectifs estimés à partir du tableau 23,
après recodification de la variable origine sociale.

"Variables explicatives"		Réussite	Échec	Total
Origine sociale	- Industriels, prof. lib.,… - Agriculteurs, ouvriers,...	1530 = 51,3 % 1312 = 42,8 %	1454 = 48,7 % 1752 = 57,2 %	2984 3064
Travail salarié	- Ne travaille pas - Travaille	2279 = 58,4 % 563 = 26,2 %	1624 = 41,6 % 1582 = 73,8 %	3903 2145
Âge à la 1ère inscription	- Moins de 20 ans - 20 ans et plus	1914 = 61,3 % 928 = 31,7 %	1206 = 38,7 % 2000 = 68,3 %	3120 2928
Mention au Bac	- Mention - Pas de mention	897 = 62,5 % 1945 = 42,2 %	539 = 37,5 % 2667 = 57,8 %	1436 4612
Section dans le secondaire	- Classique - Moderne	1147 = 63,0 % 1695 = 40,1 %	673 = 37,0 % 2533 = 59,9 %	1820 4228

Tableau 25. Influence sur la réussite au CELG
de chacune des "variables explicatives" considérées isolément.

Selon les différences des pourcentages de réussite calculées sur le tableau 25, la variable qui influe le plus sur la réussite au CELG est le fait de ne pas exercer une profession parallèlement à la poursuite des études supérieures :

- origine sociale : 51,3 % - 42,8 % = 8,5 %
- travail salarié : 58,4 % - 26,2 % = 32,2 %
- âge à la première inscription : 61,3 % - 31,7 % = 29,6 %
- mention au Bac : 62,5 % - 42,2 % = 20,3 %
- section dans le secondaire : 63,0 % - 40,1 % = 22,9 %

Quelle est maintenant la variable qui, le travail salarié mis à part, exerce le plus d'influence sur le taux de réussite ? Pour répondre à cette question, il suffit de considérer séparément les deux sous-populations homogènes du point de vue de cette variable, et d'appliquer le raisonnement précédent d'une part aux étudiants qui travaillent, d'autre part aux étudiants qui n'exercent pas de profession. Pour faciliter la présentation ultérieure de l'ensemble des résultats (dans le schéma 5 ci-après), nous affecterons à chaque sous-population issue de la segmentation un numéro indiquant le niveau de segmentation (ici : 1) et un numéro d'ordre (ici : 1 pour les étudiants salariés, et 2 pour ceux qui ne travaillent pas).

Au vu du tableau 26, pour les 2 145 étudiants exerçant une profession, la variable qui influe le plus sur la réussite est la mention au Bac (écart = 13,1 %), suivi par l'âge à la première inscription (10,8 %), la section dans le secondaire (2,7 %), et enfin l'origine sociale (1,1 %). La faible influence de l'origine sociale peut paraître en contradiction avec ce que l'on sait par ailleurs sur la sélection sociale dans le système éducatif. Nous y reviendrons plus loin, mais l'on peut d'ores et déjà faire l'hypothèse que l'origine sociale et la nécessité de travailler pendant ses études sont liées (il suffit pour s'en convaincre de constituer le tableau croisé : travail × origine sociale), et que l'effet de l'origine sociale est ici masqué par celui de l'obligation de travailler.

"Variables explicatives"		Réussite	Échec	Total
Origine sociale	- Industriels, prof. lib.,… - Agriculteurs, ouvriers,...	177 = 25,5 % 386 = 26,6 %	517 = 74,5 % 1065 = 73,4 %	694 1451
Âge à la 1ère inscription	- Moins de 20 ans - 20 ans et plus	109 = 35,5 % 454 = 24,7 %	198 = 64,5 % 1384 = 75,3 %	307 1838
Mention au Bac	- Mention - Pas de mention	176 = 36,4 % 387 = 23,3 %	307 = 63,2 % 1275 = 76,7 %	483 1662
Section dans le secondaire	- Classique - Moderne	98 = 28,5 % 465 = 25,8 %	246 = 71,5 % 1336 = 74,2 %	344 1801

Tableau 26. Influence sur la réussite au CELG
des "variables explicatives" pour les étudiants exerçant une profession (1.1).

En procédant de la même manière pour les 3 903 étudiants qui n'exercent pas de profession (tableau 27), on observe que, si les effets des "variables explicatives" sur la réussite sont généralement plus marqués, l'ordre dans lequel ils se classent est le même que pour les étudiants salariés : mention au Bac (22,9 %), âge à la première inscription (20,7 %), section dans le secondaire (20,4 %), et origine sociale (1,7 %).

"Variables explicatives"		Réussite	Échec	Total
Origine sociale	- Industriels, prof. lib.,… - Agriculteurs, ouvriers,...	1353 = 59,1 % 926 = 57,4 %	937 = 40,9 % 687 = 42,6 %	2290 1613
Âge à la 1ère inscription	- Moins de 20 ans - 20 ans et plus	1805 = 64,2 % 474 = 43,5 %	1008 = 35,8 % 616 = 56,5 %	2813 1090
Mention au Bac	- Mention - Pas de mention	721 = 75,7 % 1558 = 52,8 %	232 = 24,3 % 1392 = 47,2 %	953 2950
Section dans le secondaire	- Classique - Moderne	1049 = 71,1 % 1230 = 50,7 %	427 = 28,9 % 1197 = 49,3 %	1476 2427

Tableau 27. Influence sur la réussite au CELG
des "variables explicatives" pour les étudiants n'exerçant pas de profession (1.2).

La même démarche conduit à distinguer, au deuxième niveau de segmentation, quatre sous-populations d'étudiants. L'influence sur la réussite à l'examen, au sein de chacune de ces sous-populations, des trois "variables explicatives" restantes est l'objet du tableau 28. C'est l'âge à la première inscription qui a l'influence la plus forte, sauf pour les étudiants qui ne travaillent pas et qui n'ont pas de mention au Bac : pour ces derniers, c'est la section suivie dans l'enseignement secondaire qui influe le plus sur la réussite.

Il faut toutefois rappeler que les écarts de pourcentages ne sont pas la meilleure mesure de liaison statistique entre deux variables, et que d'autres coefficients plus appropriés (mais moins aisés à calculer) n'aboutiraient pas aux mêmes constats ; pour pouvoir affirmer que c'est bien l'âge à la première inscription qui a le plus d'influence à ce niveau, il conviendrait de vérifier que le coefficient de contingence φ ou l'indice de Belson conduisent bien au même résultat. Si l'on s'en tient à la comparaison des différences de pourcentages, on observe que le "pouvoir explicatif" des variables prises en compte est plus marqué dans certaines sous-populations (étudiants ne travaillant pas, ayant eu une mention au Bac), qui affichent des écarts importants, que dans d'autres (étudiants qui travaillent et n'ont pas eu de mention) ; ceci aura, comme nous le verrons, un effet sur la poursuite de la procédure de segmentation. Enfin, on notera que, pour les étudiants qui travaillent et ont obtenu une mention au Bac, l'influence de l'origine sociale s'exerce en sens inverse : le taux de réussite au CELG est plus élevé chez ceux issus de milieux "modestes" (38,9 %) que chez ceux issus de milieux aisés (30,1 %).

Travaillent, sans mention au Bac (2.1)					Écart
- origine sociale	(137/561 = 24,4 %)	-	(250/1101 = 22,7 %)	=	1,7 %
- âge à la $1^{\text{ère}}$ inscription	(67/234 = 28,6 %)	-	(320/1428 = 22,4 %)	=	6,2 %
- section du secondaire	(71/283 = 25,1 %)	-	(316/1379 = 22,9 %)	=	2,2 %
Travaillent, avec mention au Bac (2.2)					**Écart**
- origine sociale	(40/133 = 30,1 %)	-	(136/350 = 38,9 %)	=	- 8,8 %
- âge à la $1^{\text{ère}}$ inscription	(42/73 = 57,5 %)	-	(134/410 = 32,7 %)	=	24,8 %
- section du secondaire	(27/61 = 44,3 %)	-	(149/422 = 35,3 %)	=	9,0 %
Ne travaillent pas, sans mention au Bac (2.3)					**Écart**
- origine sociale	(936/1748 = 53,5 %)	-	(622/1202 = 51,7 %)	=	1,8 %
- âge à la $1^{\text{ère}}$ inscription	(1166/2025 = 57,6 %)	-	(392/925 = 42,4 %)	=	15,2 %
- section du secondaire	(624/987 = 63,2 %)	-	(934/1963 = 47,6 %)	=	15,6 %
Ne travaillent pas, avec mention au Bac (2.4)					**Écart**
- origine sociale	(417/542 = 76,9 %)	-	(304/411 = 74,0 %)	=	5,6 %
- âge à la $1^{\text{ère}}$ inscription	(639/788 = 81,1 %)	-	(82/165 = 49,7 %)	=	31,4 %
- section du secondaire	(425/489 = 86,9 %)	-	(296/464 = 63,8 %)	=	23,1 %

Tableau 28. Influence sur la réussite au CELG des "variables explicatives"
pour les quatre sous-populations d'étudiants distinguées au deuxième niveau de segmentation.

On peut continuer à appliquer la même démarche tant qu'il reste des "variables explicatives" qui n'ont pas été utilisées dans la segmentation de la population étudiée (le tableau 29 correspond au troisième niveau de segmentation). Quel que soit le coefficient utilisé pour mesurer l'influence de ces "variables explicatives", il est conseillé d'interrompre le processus de segmentation dès que la liaison mesurée est trop faible, ou la sous-population segmentée trop petite. Nous avons ici fixé deux seuils d'arrêt de la démarche : ne pas segmenter de sous-populations d'une taille inférieure à 100 individus, et ne pas prendre en considération les écarts de pourcentages inférieurs à 5 %. Ce dernier critère justifie l'arrêt de la segmentation pour les étudiants qui travaillent et n'ont pas eu de mention au Bac (sous-populations 3.1 et 3.2, dont les écarts sont compris entre 2,6 % et 2,0 %). Quant aux étudiants salariés ayant obtenu une mention au Bac et inscrits à moins de 20 ans (sous-population 3.4), c'est sa petite taille (73 individus) qui explique l'arrêt de la segmentation, malgré une influence importante de la section suivie dans le secondaire (écart = 23,1 %).

Travaillent, sans mention au Bac, inscrits à 20 ans et plus (3.1)					Écart
- origine sociale	(121/508 = 23,8 %)	-	(199/920 = 21,6 %)	=	2,2 %
- section du secondaire	(59/244 = 24,2 %)	-	(261/1184 = 22,0 %)	=	2,2 %
Travaillent, sans mention au Bac, inscrits à moins de 20 ans (3.2)					**Écart**
- origine sociale	(16/53 = 30,2 %)	-	(51/181 = 28,2 %)	=	2,0 %
- section du secondaire	(12/39 = 30,8 %)	-	(55/195 = 28,2 %)	=	2,6 %
Travaillent, avec mention au Bac, inscrits à 20 ans et plus (3.3)					**Écart**
- origine sociale	(35/123 = 28,5 %)	-	(99/287 = 34,5 %)	=	- 6,0 %
- section du secondaire	(20/52 = 38,5 %)	-	(114/358 = 31,8 %)	=	6,7 %
Travaillent, avec mention au Bac, inscrits à moins de 20 ans (3.4)					**Écart**
- origine sociale	(5/10 = 50,0 %)	-	(37/63 = 58,7 %)	=	- 8,7 %
- section du secondaire	(7/9 = 77,8 %)	-	(35/64 = 54,7 %)	=	23,1 %
Ne travaillent pas, sans mention au Bac, section moderne (3.5)					**Écart**
- origine sociale	(508/1081 = 47,0 %)	-	(426/882 = 48,3 %)	=	- 1,3 %
- âge à la 1ère inscription	(647/1257 = 51,5 %)	-	(287/706 = 40,7 %)	=	10,8 %
Ne travaillent pas, sans mention au Bac, section classique (3.6)					**Écart**
- origine sociale	(428/667 = 64,2 %)	-	(196/320 = 61,3 %)	=	2,9 %
- âge à la 1ère inscription	(519/768 = 67,6 %)	-	(105/219 = 47,9 %)	=	19,7 %
Ne travaillent pas, avec mention au Bac, inscrits à 20 ans et plus (3.7)					**Écart**
- origine sociale	(45/84 = 53,6 %)	-	(37/81 = 45,7 %)	=	7,9 %
- section du secondaire	(30/42 = 71,4 %)	-	(52/123 = 42,3 %)	=	29,1 %
Ne travaillent pas, avec mention au Bac, inscrits à moins de 20 ans (3.8)					**Écart**
- origine sociale	(372/458 = 81,2 %)	-	(267/330 = 80,9 %)	=	0,3 %
- section du secondaire	(395/447 = 88,4 %)	-	(244/341 = 71,6 %)	=	16,8 %

Tableau 29. Influence sur la réussite au CELG des "variables explicatives" pour les huit sous-populations d'étudiants distinguées au troisième niveau de segmentation.

Au dernier niveau de segmentation, il ne reste plus qu'une seule "variable explicative". Il suffit, pour effectuer cette dernière étape, d'identifier dans le tableau 24 les deux lignes qui correspondent à la sous-population cherchée, et de calculer la différence des pourcentages de réussite. En appliquant les critères d'arrêt que nous avons fixés, on voit que seules trois sous-populations ont à la fois un effectif assez important et un écart de pourcentages suffisant pour être encore segmentées. On observe également que la "variable explicative" non encore utilisée est, dans tous les cas, l'origine sociale des étudiants.

c) Comment interpréter ces résultats ?

Le résultat de la segmentation est représenté dans le schéma 5. La case en haut du schéma (0.1) représente l'ensemble de la population étudiée : 6 048 étudiants, avec un taux de réussite de 47,0 %. Au premier niveau de segmentation, le facteur qui exerce la plus forte influence sur la réussite au CELG distingue les étudiants exerçant une profession (1.1), dont le taux de réussite n'est que de 26,2 %, et ceux qui ne travaillent pas (1.2), qui réussissent à 58,4 %. À chaque nœud de l'arborescence, la sous-population au plus fort taux de réussite se trouve placée à droite. On remarquera toutefois que cette règle de présentation n'implique pas qu'à un même niveau, les sous-populations soient rangées dans un ordre croissant de taux de réussite (voir par exemple le troisième niveau). Par contre, la sous-population la plus à gauche du schéma (3.1) est bien celle dont le taux de réussite est le plus faible (22,4 %), et celle placée à l'extrême droite (4.16) est celle qui présente le taux le plus élevé (88,4 %).

Que nous apprend la lecture de ce schéma ? Tout d'abord que, si l'on avait voulu pronostiquer la réussite d'un étudiant parisien au CELG en 1963, le meilleur critère est le fait que l'étudiant exerce ou non une profession parallèlement à la poursuite de ses études. L'information la plus importante ensuite pour un tel pronostic aurait été la mention au Bac. Puis, selon les cas, c'est l'âge à la première inscription dans l'enseignement supérieur ou la section suivie dans l'enseignement secondaire qui aurait permis d'affiner le pronostic. L'origine sociale intervient peu pour le pronostic ; en outre, lorsqu'elle intervient, son influence ne s'exerce pas toujours dans le même sens. Par exemple, pour les étudiants ne travaillant pas, ayant eu une mention au Bac, entrés dans l'enseignement supérieur à moins de 20 ans, et ayant suivi une section moderne (4.15), le taux de réussite est plus élevé chez ceux dont l'origine sociale est la plus modeste (sous-population 5.30, à comparer avec 5.29). Dans le cas général, c'est la relation inverse qui s'observe (segmentation des sous-populations 4.11 et 4.13).

Il aurait été pratiquement impossible de dégager ces diverses influences, et en particulier ces interactions, à partir d'une simple lecture du tableau 23. Il reste à interpréter ces observations. Il est assez facile, pour tout esprit entraîné au maniement des mots et des concepts (c'est l'objectif de l'exercice scolaire appelé "dissertation"), de concevoir une explication théorique applicable à chaque cas particulier ("explication ad hoc"). Lorsque les étudiants de milieu aisé réussissent mieux que les autres, il est "évident" que cela tient au niveau culturel des parents et à leur aide financière ; lorsqu'ils réussissent moins bien, il est non moins "évident" qu'ils n'ont pas été soumis, comme les étudiants d'origine modeste, à une sélection sévère dès l'enseignement secondaire…Comme le dit Paul Lazarsfeld, "ce qui est évident, c'est que quelque chose ne va pas dans tout ce raisonnement sur l'évidence. En réalité, il faudrait le retourner : puisque toute espèce de réaction humaine est concevable, il est d'une grande importance de savoir quelles réactions se produisent, en fait, le plus fréquemment et dans quelles conditions" [1]. C'est pourquoi il faut commencer par s'interroger sur la

[1] Cité page 142 de Pierre Bourdieu, Jean-Claude Chamboredon, Jean-Claude Passeron, *Le métier de sociologue*, Paris, Mouton-Bordas, 1968.

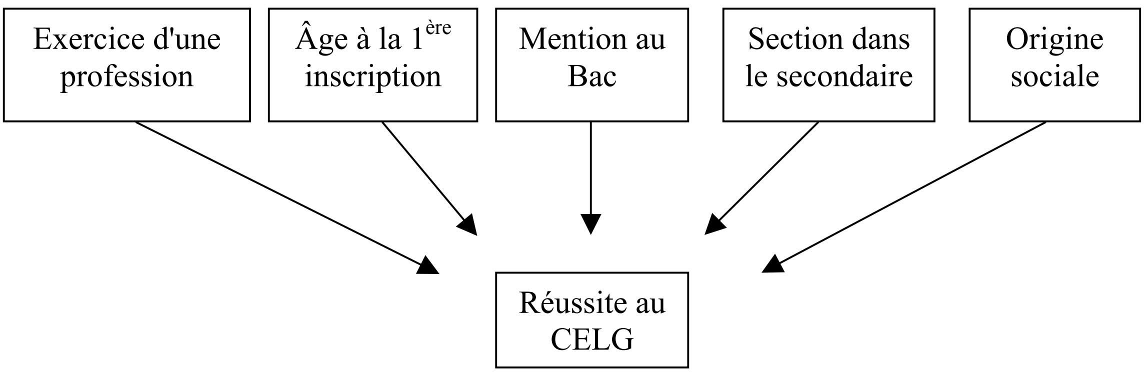
signification des données dont nous disposons, et sur le schéma explicatif qui sous-tend la démarche utilisée.

La population étudiée est la totalité des étudiants inscrits en première année de la Faculté des Lettres et Sciences Humaines de Paris (Sorbonne) pour l'année 1962-63. Par rapport aux autres filières d'études supérieures, cette "propédeutique littéraire" est un prolongement des études secondaires : aucune des matières étudiées n'est nouvelle, seul (en principe) le niveau diffère. L'examen final ne comporte que trois épreuves écrites, portant (selon les options choisies) sur le français, la philosophie, l'histoire, la géographie, les langues (vivantes ou mortes). Les enseignements sont peu nombreux, et l'assiduité y est peu contrôlée. Il est par conséquent possible de préparer le Certificat d'Études Littéraires Générales en exerçant une profession, ou en poursuivant d'autres études plus contraignantes.

Les données sont tirées des fiches d'inscription et des listes de résultats aux examens établis par la Faculté. Ce sont au départ des données exhaustives, bien que 12,6 % des 6 919 étudiants recensés aient dû être éliminés du tableau 23 en raison d'informations incomplètes. On ignore naturellement si les 871 étudiants exclus du tableau 23 sont différents des 6 048 étudiants qui y figurent ; mais on peut faire l'hypothèse que les dossiers incomplets sont plus fréquents chez les étudiants les plus défavorisés (hypothèse du "cumul des handicaps"). Il faudra donc garder présent à l'esprit que les observations analysées ne sont plus exhaustives, et risquent de présenter un biais non identifié (et donc de n'être plus strictement représentatives de la population étudiée).

Le taux de réussite a été calculé sur le nombre d'inscrits en début d'année, et non sur le nombre de présents à l'examen. Or, on sait qu'il y a eu 34 % d'abandons en cours d'année [1]. Il est clair que la signification des "non réussites" n'est pas la même selon qu'il s'agit d'un renoncement (provisoire ?) à ce type d'études ou d'une préparation insuffisante à l'examen ; d'ailleurs, le taux d'abandon est sensiblement plus élevé chez les étudiants d'origine modeste (ibid.). En outre, les abandons peuvent être dus également au peu de débouchés professionnels correspondant à ce type d'études littéraires.

Enfin, la démarche d'analyse (inspirée des méthodes de segmentation) que nous avons suivie repose sur un schéma théorique particulièrement simpliste : il y a une "variable à expliquer" (le taux de réussite au CELG), et cinq "variables explicatives" supposées agir directement sur la "variable à expliquer".



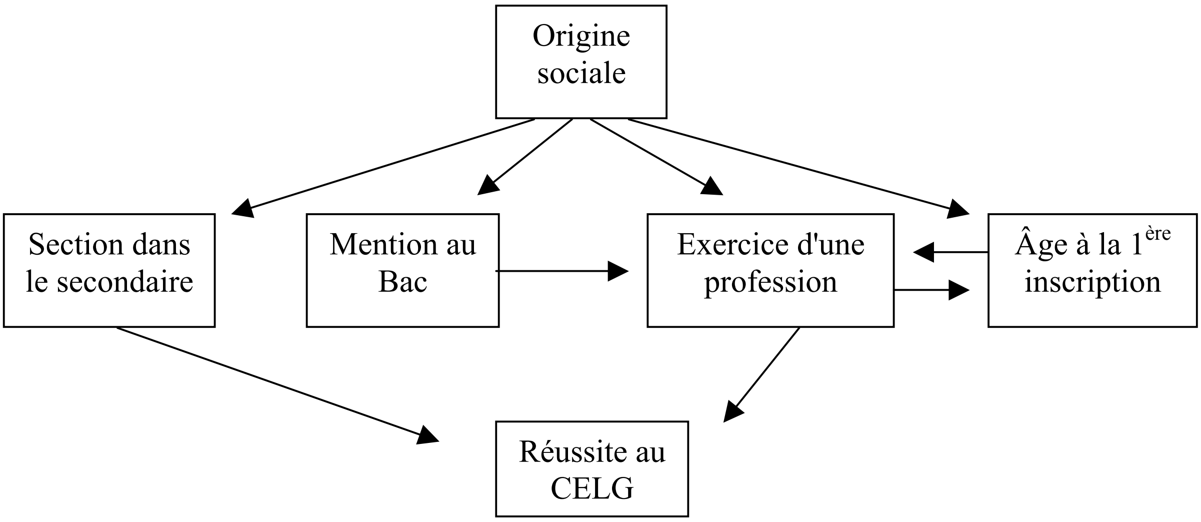
[1] Noëlle Bisseret, *op. cit.*, page 189.

Appliqué au problème de la réussite universitaire, un tel schéma a peu de chances d'être considéré comme un schéma explicatif valide. Par contre, il constitue un excellent schéma prévisionnel. Si notre objectif n'est pas d'expliquer les échecs et les réussites, mais plus modestement de pronostiquer l'issue de la première année d'études littéraires à partir des informations dont on dispose sur les étudiants, le schéma 5 présente un grand intérêt pratique : il permet de distinguer entre des taux de réussite s'échelonnant de 22,4 % à 88,4 %, et ceci à partir de seulement cinq informations dichotomiques. Si l'on admet que, d'une année sur l'autre, ces résultats varieront peu, cette démarche a donc une valeur pronostique certaine (elle peut servir de base, par exemple, pour l'attribution d'aides matérielles ou de soutien technique aux étudiants les plus exposés aux échecs). En outre, sur le plan théorique, ces résultats révèlent d'importants effets d'interaction.

d) Comment construire un schéma explicatif ?

Telle que nous l'avons présentée, cette démarche se justifie pleinement si nous ne disposons par ailleurs d'aucune autre information sur le phénomène étudié ; elle constitue alors une excellente procédure d'exploration des relations statistiques entre les variables. Dans ce cas précis, il est évident que nous disposons d'autres informations, et que nous pouvons élaborer un schéma plus réaliste, même en nous limitant aux cinq variables du tableau 23. Par exemple, les travaux d'Alain Girard sur la sélection sociale dans le système éducatif ont montré que l'âge à l'entrée dans une filière d'enseignement jouait un rôle important, et que celui-ci, tout comme les résultats scolaires, était fortement lié à l'origine sociale. On peut donc raisonnablement postuler que l'origine sociale influe directement sur l'âge à la première inscription, le choix de la section suivie dans l'enseignement secondaire, la mention obtenue au Bac, et la nécessité de travailler pour poursuivre ses études.

On peut ensuite faire l'hypothèse que le fait d'avoir eu une mention au Bac et d'être relativement jeune au moment de l'entrée en Faculté augmentent les chances d'obtenir une aide (bourse, prêt d'honneur) pour la poursuite des études dans de meilleures conditions ; ces deux variables influeraient donc directement sur l'obligation de travailler. En outre, l'âge à la première inscription peut également être plus tardif lorsqu'il s'agit de salariés reprenant des études en continuant de travailler : on a donc aussi une influence possible de l'exercice d'une profession sur l'âge à la première inscription. Enfin, on pourrait supposer que ce sont l'obligation d'exercer ou non une profession et la possibilité de bénéficier ou non des acquis d'une formation classique qui influent principalement sur la réussite dans les études littéraires.



Le tableau 30 (constitué comme précédemment à partir du tableau 24) vise à mesurer la force de ces influences directes. On y constate que l'origine sociale a une influence très forte sur la nécessité de travailler pendant la poursuite des études supérieures ; et une influence moindre, mais réelle, sur le choix de la filière dans le secondaire, et sur l'âge à l'entrée dans l'enseignement supérieur. Par contre, elle ne semble pas ici avoir d'influence sensible sur la mention au Bac (on ne peut naturellement pas en déduire que l'origine sociale n'a pas d'influence sur les résultats au Bac, puisque ces données ne portent pas sur la totalité des bacheliers). Ces constats recoupent des observations courantes en sociologie de l'éducation : plus fréquemment que les autres étudiants, ceux d'origine modeste ont suivi des sections "modernes", accèdent relativement tard à l'enseignement supérieur, sont privés de l'aide de leur famille (ce qui les oblige à travailler) ; en outre, il peut arriver qu'ils soient sur-sélectionnés par rapport aux étudiants de milieux aisés (d'où le taux de mentions au Bac).

	Origine sociale		Écart
	"aisée"	"modeste"	
- 1ère inscription à < 20 ans	1711/2984 = 57,3 %	1409/3064 = 46,0 %	11,3 %
- section classique	1143/2984 = 38,3 %	677/3064 = 22,1 %	16,2 %
- mention au Bac	675/2984 = 22,6 %	761/3064 = 24,8 %	- 2,2 %
- exercice d'une profession	694/2984 = 23,3 %	1451/3064 = 47,4 %	- 24,1 %

Tableau 30. Influence de l'origine sociale sur les autres "variables explicatives".

% exerçant une profession	1ère inscription à ≥ 20 ans	1ère inscription à < 20 ans	Écart
- origine "aisée" - origine "modeste"	631/1273 = 49,6 % 1207/1655 = 72,9 %	63/1711 = 3,7 % 244/1409 = 17,3 %	45,9 % 55,6 %
- ensemble	1838/2928 = 62,8 %	307/3120 = 9,8 %	53,0 %
% exerçant une profession	**Pas de mention au Bac**	**Mention au Bac**	**Écart**
- origine "aisée" - origine "modeste"	561/2309 = 24,3 % 1101/2303 = 47,8 %	133/675 = 19,7 % 350/761 = 46,0 %	4,6 % 1,8 %
- ensemble	1662/4612 = 36,0 %	483/1436 = 33,6 %	2,4 %

Tableau 31. Influence de l'âge à la première inscription et de la mention au Bac sur l'exercice d'une profession.

Alors que la mention au Bac semble n'avoir qu'un effet très limité sur le fait que l'étudiant travaille, l'âge à l'entrée dans l'enseignement supérieur et l'exercice d'une profession sont fortement liées, surtout chez les étudiants d'origine "modeste" (tableau 31). Faute d'informations complémentaires, en particulier sur le temps écoulé entre l'obtention du Bac et le début des études supérieures, on ne peut ici déterminer dans quel sens l'influence s'exerce : est-on en présence de personnes actives ayant choisi d'entreprendre des études supérieures après avoir commencé à travailler, ou d'étudiants ne parvenant à vivre sans une source de revenus tirés d'une activité rémunérée ?

Cette analyse pourrait être poussée plus loin, en particulier en s'inspirant de la démarche de l'analyse causale (*path analysis*). Remarquons pour conclure qu'il n'est pas nécessaire d'appliquer aveuglément la procédure de segmentation en commençant par la "variable explicative" ayant la plus forte influence sur la "variable à expliquer". Rien n'interdit au sociologue de commencer la segmentation par la variable qu'il estime être à la source du phénomène étudié, et de donner par exemple à l'origine sociale la place prépondérante. À titre d'exercice, on pourrait refaire le schéma 5 en commençant par segmenter la population selon l'origine sociale, et en choisissant ensuite, à chaque niveau de segmentation, la variable la plus intéressante du point de vue sociologique.

3.3. Récapitulation et compléments.

Après avoir rappelé comment aborder un tableau multidimensionnel, nous revenons dans ce sous-chapitre sur les notions de relation statistique (en introduisant celle de *liaison partielle*), et de causalité.

a) Comment lire un tableau à plus de deux dimensions.

Dès que plusieurs variables sont impliquées dans un tableau statistique, on ne peut guère se passer d'hypothèses de type causal. Nous l'avons vu pour les tableaux à double entrée : étudier les liaisons statistiques (les variations concomitantes) entre deux variables présuppose le plus souvent, dans l'esprit du chercheur, que l'une des variables pourrait avoir une influence sur l'autre. Lorsque le tableau est plus complexe, il n'est plus possible de se passer d'une telle présupposition : décrire, c'est alors tenter d'expliquer, et il serait vain de vouloir extraire des informations d'une distribution multidimensionnelle sans avoir défini au préalable une problématique.

Le schéma explicatif le plus simple consiste à distinguer, parmi les variables du tableau, une "variable à expliquer", c'est à dire une variable qui : 1) présente pour le sociologue un grand intérêt théorique (il s'agit d'une caractéristique, d'un comportement, d'une opinion, d'une intention ou d'une attitude qui se trouve au centre de sa problématique) ; 2) cette "variable à expliquer" ne peut être qu'une conséquence éventuelle de certaines des autres variables de la distribution, car ces "variables explicatives" potentielles lui sont antérieures. Un tel schéma correspond à une situation dans laquelle on ne dispose d'aucune autre information sur les variables du tableau (par exemple parce qu'il s'agit d'un domaine nouveau, ou encore peu exploré), ou bien dans laquelle on décide de négliger ces informations (par exemple en raison de la trop grande complexité du schéma qui en résulterait, alors que l'objectif visé serait seulement la décision ou la prévision à court terme).

Les informations que l'on extrait d'un tableau complexe sont d'autant plus riches que le schéma explicatif que l'on a construit est plus élaboré. Naturellement, ce schéma ne peut se fonder que sur l'expérience : expérience personnelle du lecteur (connaissance du terrain), et expérience des chercheurs qui ont déjà travaillé sur le sujet ou des sujets connexes (tirée de la littérature scientifique). C'est le modèle explicatif qui dicte la démarche à suivre pour analyser le tableau : la règle en la matière étant que l'on n'obtient des informations utilisables que si l'on se pose des questions. Il convient donc d'"interroger les données" avec clarté et précision. Pour cela, chaque fois que l'on postulera une relation causale entre deux variables, on cherchera s'il existe une liaison statistique entre ces variables.

Les schémas que nous avons examinés étaient particulièrement simples, puisqu'il n'y avait qu'une seule "variable à expliquer". Dans le premier exemple (à trois variables), il n'y avait guère que quatre problématiques possibles : 1) les trois variables n'ont aucune relation causale entre elles ; 2) la classe objective seule influe sur le sentiment d'appartenance ; 3) la classe subjective seule influe sur le sentiment d'appartenance ; 4) les deux influent sur ce sentiment. Dans ce dernier cas, il fallait en outre rechercher s'il existait des effets d'interaction. Dans le second exemple (à six variables), il y aurait eu 31 schémas simples possibles [1], entre lesquels il aurait fallu choisir. Naturellement, si, à ces influences sur la "variable à expliquer", l'on ajoute des relations entre les "variables explicatives", le nombre de schémas explicatifs possibles croît très rapidement en fonction du nombre de variables ; on ne

[1] Pour *n* "variables explicatives" et une seule "variable à expliquer", il y aurait $2^n - 1$ schémas simples possibles (sans tenir compte des influences des "variables explicatives" les unes sur les autres).

peut donc raisonnablement les examiner tous. C'est dire l'importance qu'il y a de disposer d'un ensemble d'hypothèses bien structurées, même pour une simple lecture de tableau statistique.

b) Liaisons globales et liaisons partielles.

Nous avons vu qu'il y a une relation statistique (liaison, association, contingence, corrélation) entre deux variables lorsqu'un état donné de l'une des variables se trouve associé plus fréquemment qu'en moyenne à un état donné de l'autre variable ; on détecte facilement les variations concomitantes (les *covariations*) au fait que les pourcentages de ligne ou les pourcentages de colonne ne sont pas semblables à ceux des marges correspondantes.

Lorsque plus de deux variables sont impliquées, deux nouvelles notions apparaissent : celle d'*interaction*, et celle de *liaison partielle*. Une liaison partielle (corrélation partielle) est celle que l'on observe entre deux variables lorsque l'on a préalablement annulé l'influence d'autres variables en les maintenant constantes. Cette démarche prend sa source dans les sciences expérimentales, qui raisonnent généralement "toutes choses égales par ailleurs" [1]. L'avantage de l'expérimentation est de pouvoir faire varier une "variable explicative" en maintenant fixes toutes les autres, et d'isoler ainsi les effets spécifiques de cette variable sur la "variable à expliquer". Les sociologues ont rarement cette possibilité ; mais ils peuvent s'inspirer de la méthode expérimentale en commençant par isoler les sous-populations homogènes selon les variables que l'on prétend maintenir constantes ; puis en mesurant la liaison entre deux variables sur chacune de ces sous-populations ; et enfin en calculant la moyenne de ces mesures (pondérée par les effectifs de chaque sous-population).

Ainsi, dans le premier exemple, nous aurions pu chercher à mesurer l'influence, sur le sentiment d'appartenir à la classe ouvrière, du seul fait d'être ou non ouvrier (classe objective) ; le tableau 32 oppose les données d'observation (calculées à partir du tableau 17) et les effectifs "théoriques" correspondant à l'indépendance des variables [2]. Il donne ainsi tous les éléments du calcul du coefficient d'association φ ("phi"), tiré du χ^2 ("khi deux"). Ce coefficient, qui varie de 0 à 1, est d'autant plus élevé que la distribution est plus éloignée de l'hypothèse d'indépendance. La liaison ainsi mesurée ($\varphi = 0{,}3485$) est la *liaison globale* entre les deux variables.

---

[1] Sur la démarche expérimentale dans les sciences sociales, on lira : Benjamin Matalon, *Décrire, expliquer, prévoir*, Paris, Armand Colin, 1988.
[2] Afin d'augmenter la précision des calculs du χ^2, nous avons conservé une décimale aux effectifs correspondant à l'hypothèse d'indépendance. S'agissant d'effectifs "théoriques" (c'est à dire : non réels), il n'est pas nécessaire que ce soient des nombres entiers.

Sentiment d'appartenance	Classe objective ouvrière	Classe objective autre	Total
- classe ouvrière	135	78	213
- autres réponses	132	365	497
Total	267	443	710

Données observées

Sentiment d'appartenance	Classe objective ouvrière	Classe objective autre	Total
- classe ouvrière	80,1	132,9	213,0
- autres réponses	186,9	310,1	497,0
Total	267,0	443,0	710,0

Hypothèse d'indépendance

Tableau 32. Mesure de la liaison globale entre la classe d'appartenance et le sentiment d'appartenance (χ^2 = 86 ; φ = 0,35).

Origine ouvrière			
Sentiment d'appartenance	Classe objective ouvrière	Classe objective autre	Total
- classe ouvrière	84	33	117
- autres réponses	74	84	158
Total	158	117	275
Autre origine			
Sentiment d'appartenance	Classe objective ouvrière	Classe objective autre	Total
- classe ouvrière	51	45	96
- autres réponses	58	281	339
Total	109	326	435

Données observées

Origine ouvrière			
Sentiment d'appartenance	Classe objective ouvrière	Classe objective autre	Total
- classe ouvrière	67,2	49,8	117,0
- autres réponses	90,8	67,2	158,0
Total	158,0	117,0	275,0
Autre origine			
Sentiment d'appartenance	Classe objective ouvrière	Classe objective autre	Total
- classe ouvrière	24,1	71,9	96,0
- autres réponses	84,9	254,1	339,0
Total	109,0	326,0	435,0

Hypothèse d'indépendance

Tableau 33. Mesure de la liaison partielle entre la classe d'appartenance et le sentiment d'appartenance à origine sociale constante (χ^2 = 69 ; φ = 0,31).

On peut chercher à mesurer de la même manière l'influence, sur ce sentiment d'appartenance, du fait d'être ouvrier, *moins* l'influence de l'origine sociale : on sait en effet que l'origine sociale influe à la fois sur le fait d'être ouvrier (classe objective) et sur le sentiment d'appartenance (classe subjective). Le tableau 33 illustre le moyen de "maintenir constante" la variable origine sociale : on commence par considérer séparément les deux sous-populations homogènes selon cette variable. On peut naturellement calculer un coefficient d'association pour chacune de ces sous-populations prises isolément ; on s'aperçoit, à cette occasion, que la liaison entre classe objective et sentiment d'appartenance est plus faible chez les hommes d'origine ouvrière ($\chi^2 = 17$, $\varphi = 0,25$) que chez ceux qui n'ont pas un père ouvrier ($\chi^2 = 51$, $\varphi = 0,34$).

On peut également calculer un coefficient d'association *partielle* (c'est à dire ici : à origine sociale constante) à partir de l'ensemble des distributions du tableau 33 ($\chi^2 = 69$; $\varphi = 0,31$). C'est la solution la plus couramment adoptée. Les coefficients de *corrélation partielle* découlent du même type de raisonnement, appliqué aux variables mesurables.

c) Les interactions, et les liaisons cachées.

Les interactions entre deux variables ou plus traduisent l'influence *conjointe* que celles-ci exercent sur d'autres variables. Dans le premier exemple, nous avons mis cet effet en évidence à partir de la décomposition des pourcentages ; dans le second exemple (schéma 5), nous avons qualifié d'effet d'interaction le fait que, par exemple, une origine sociale "modeste" pouvait, selon les cas, entraîner l'échec (segmentation des sous-populations 4.11 et 4.13), ou au contraire favoriser la réussite (segmentation de la sous-population 4.15). Le lien entre ces deux aspects du même phénomène apparaîtra dans l'exemple ci-après (fictif, mais inspiré d'un cas réel). Cet exemple, dû à Sonquist et Morgan, sera de plus l'occasion de montrer qu'une liaison entre deux variables peut être masquée par des effets d'interaction.

Supposons une population de 1 000 patients ayant été hospitalisées, sur lesquels on étudie la durée totale des périodes d'hospitalisation ("variable à expliquer") à partir de leur âge et de leur sexe ("variables explicatives"). L'âge et la durée d'hospitalisation ont été dichotomisés à partir de la médiane. Les tableaux à deux dimensions croisant les "variables explicatives" avec la "variable à expliquer" montrent que ni l'âge, ni le sexe n'ont de liaison directe avec la durée d'hospitalisation (tableaux 34 et 35). À titre de contrôle, on constate que l'âge et le sexe n'ont entre eux aucune liaison statistique (le contenu du tableau qui les croise est identique à celui des tableaux 34 et 35 ; il correspond par conséquent au cas de l'indépendance des variables).

Durée d'hospitalisation	Âge		Total
	Jeunes	Âgés	
- longue	250	250	500
- courte	250	250	500
Total	500	500	1 000

Tableau 34. Influence de l'âge sur la durée d'hospitalisation.

Durée d'hospitalisation	Sexe		Total
	H	F	
- longue	250	250	500
- courte	250	250	500
Total	500	500	1 000

Tableau 35. Influence du sexe sur la durée d'hospitalisation.

Au vu de ces tableaux, on serait tenté d'affirmer qu'il n'existe aucune relation statistique entre les "variables explicatives" et la "variable à expliquer" ; et que par conséquent l'âge et le sexe n'ont aucune influence sur la durée d'hospitalisation. Si, par acquit de conscience, on construit le tableau croisant ces trois variables (Tableau 36), on découvre alors qu'au contraire, ces deux variables influent fortement sur la durée d'hospitalisation. Mais ici, les influences s'exercent dans des sens opposés et, finalement, se compensent : ce sont surtout les femmes jeunes et les hommes âgés qui séjournent le plus longtemps à l'hôpital. Les influences *simultanées* du sexe et de l'âge se réduisent au seul effet d'interaction, particulièrement fort ici.

Durée d'hospitalisation	Hommes		Femmes		Total
	Jeunes	Âgés	Jeunes	Âgées	
- longue	50	200	200	50	500
- courte	200	50	50	200	500
Total	250	250	250	250	1 000

Tableau 36. Influence simultanée du sexe et de l'âge sur la durée d'hospitalisation.

Dans cet exemple, les liaisons partielles, entre le sexe et la durée d'hospitalisation à âge constant, et entre l'âge et la durée d'hospitalisation à sexe constant, sont de même intensité, mais *de sens contraire*. C'est ce qui explique que les liaisons directes (entre chacune des variables explicatives et la durée d'hospitalisation) soient nulles, que l'effet d'interaction soit aussi important, et que la liaison très forte entre les "variables explicatives" considérées simultanément, et la durée d'hospitalisation ait pu être masquée.

d) Covariations et causalité.

Si la lecture d'un tableau à trois dimensions ou plus demande pratiquement toujours que l'on fasse des hypothèses causales, cela n'implique pas que les relations statistiques que l'on observe entre les variables soient la *preuve* de l'existence de relations causales. Le passage de la description à l'explication est un "saut épistémologique" important, qui engage la responsabilité du sociologue. Pour analyser un tableau, ou pour simplement décrire son contenu, on est obligé de passer par l'intermédiaire du langage, et donc des concepts qu'il exprime (les mots sont déjà des abstractions). Même à un niveau modeste, lorsque l'on se contente de décrire ce qu'un tableau nous apprend, on recourt implicitement à un modèle de la réalité que ce tableau vise à représenter. Ce modèle intègre nécessairement les théorisations que les données nous imposent : définition du domaine de recherche, choix et définition des variables, sélection de la population étudiée, préférence pour certaines méthodes de recueil, etc. (on remarquera à cette occasion que les "données" ne sont donc pas *données*, mais *construites*). Pour lire un tableau, on complète ce modèle en y intégrant ses propres hypothèses ; dans la quasi-totalité des cas, ce sont des hypothèses causales.

Pour *ajuster* ce modèle aux données du tableau, on va donc, à chaque relation causale hypothétique, essayer de faire correspondre une relation statistique (liaison, corrélation). On remaniera le modèle si nécessaire, en fonction de l'existence ou de l'absence d'une telle relation, jusqu'à ce que l'on ait obtenu un modèle relativement satisfaisant. Au terme de ce va-et-vient entre les données d'observation et leur représentation abstraite, on pourra proposer une lecture cohérente des données, dans laquelle chaque influence supposée d'une variable sur une autre est corroborée par une relation statistique observée entre deux dimensions du tableau.

Lorsque l'on aborde un tableau particulièrement complexe, comme d'ailleurs lorsqu'on lit en parallèle plusieurs tableaux sur les mêmes données ou sur le même thème, on travaille simultanément sur plusieurs variables. On a donc *a priori* le choix entre un très grand nombre de modèles possibles, qu'il serait extrêmement fastidieux, et sans aucun intérêt sociologique, d'essayer tous. Heureusement, tous ne sont pas également plausibles ; en outre, les modèles explicatifs les plus plausibles ne conviennent pas tous à chacun d'entre nous : notre expérience du terrain, notre culture en sciences sociales, nos préférences idéologiques, déterminent nos premiers choix. Nous les modulerons ensuite au fur et à mesure de leur confrontation aux données.

Le choix de l'explication définitive sera dicté par le "pouvoir explicatif" de celle-ci ; ce "pouvoir explicatif" dépend de deux facteurs : l'adéquation du modèle aux données, et les perspectives plus générales qu'il ouvre. L'adéquation aux données est une condition minimale : si l'explication est contredite par les faits, elle n'a évidemment aucun sens. Toutefois, comme les données ne sont jamais parfaites (variables importantes omises, catégories non pertinentes, échantillon peu représentatif, etc.), on peut éventuellement s'en écarter, à condition de toujours le justifier. L'ouverture de perspectives plus générales mesure en quelque sorte la fécondité du modèle explicatif. Lorsqu'il se limite à une bonne adéquation aux données, c'est un "modèle *ad hoc*", utile pour la décision ou la prévision à court terme, mais à peine plus élaboré qu'une description fidèle. Si par contre ce même modèle peut être appliqué aussi à d'autres données (s'il est *généralisable*), son intérêt théorique augmente. Le caractère généralisable d'un modèle explicatif peut venir simplement de ce qu'il s'intègre dans une théorie plus large, qui a déjà fait ses preuves ; il peut également tenir à ce que le modèle pourrait, sans grandes modifications, être appliqué à d'autres domaines des sciences sociales (c'est le cas des grands *paradigmes* de la sociologie).

Pour conclure, il est vrai que, s'il y a une influence *directe* d'une variable sur une autre (relation causale), il y a nécessairement une relation statistique entre ces variables. Mais la réciproque n'est pas vraie. Une relation statistique entre deux variables n'implique pas nécessairement une relation causale : il peut par exemple y avoir une troisième variable qui agit simultanément sur les deux autres et entraîne leur covariation (d'autres schémas plus complexes sont évidemment possibles) [1]. Par ailleurs, l'absence de relation statistique entre deux variables ne signifie pas qu'il n'existe pas de relation causale entre elles : nous avons vu qu'une forte influence peut être masquée par des effets d'interaction, ou par des variables intermédiaires. Enfin, les schémas explicatifs que nous avons proposés sont relativement simplistes : ils ne prennent pas en compte par exemple les relations de causalité circulaires, dans lesquels deux variables influent l'une sur l'autre. Ces schémas théoriques, parfois plus proches des faits observés, et d'un "pouvoir explicatif" plus grand que les modèles "mécanistes", sont d'un maniement très délicat ; ils ressortissent plus aux techniques de simulation qu'aux statistiques descriptives élémentaires.

---

[1] On donne parfois à ces relations statistiques sans relation causale directe le nom de "corrélations fallacieuses" (*spurious correlations*). C'est évidemment une faute de logique, puisque ces corrélations existent réellement, et qu'une réalité ne peut être fallacieuse ; c'est l'hypothèse d'une relation causale directe qui est ici fallacieuse.

Conclusions : Règles pratiques pour la lecture des tableaux statistiques.

Nous commençons par récapituler les principales règles empiriques que nous avons énoncées à l'occasion des exemples étudiés ; ces règles, tirées de l'expérience, permettent d'éviter les gros contresens et les interprétations approximatives dans la lecture des tableaux. Nous proposons ensuite une règle de base, qui n'est plus imposée directement par l'expérience, mais résulte d'un choix épistémologique explicite. Nous donnons enfin quelques indications bibliographiques "pour en savoir plus".

a) Règles pratiques pour la lecture des tableaux statistiques.

Règle 1 : *Définir préalablement ce dont il s'agit.*

Pour cela, à partir des commentaires qui accompagnent le tableau, replacer la collecte des données dans son contexte (retracer la genèse de la recherche). S'il s'agit d'une recherche *sociographique*, c'est à dire si l'objectif de la collecte est surtout descriptif, comme c'est le cas de la plupart des enquêtes de l'INSEE : reprendre la définition des catégories et des variables, puis analyser les concepts utilisés en explicitant les hypothèses sociologiques auxquelles ils se rattachent (par exemple : société homogène, société stratifiée, société de classes). S'il s'agit d'une recherche à visée plus *théorique* (recherche "fondamentale"), comme une thèse ou une publication du CNRS : expliciter la problématique et les références théoriques des auteurs, et définir les hypothèses retenues et les hypothèses écartées *a priori* ; dans le cas d'une enquête extensive, on peut pour cela partir de la nature et de la formulation des questions posées (par opposition à celles que l'on aurait pu poser, et que l'on a omises), et de la liste des réponses soumises aux répondants. S'il s'agit enfin d'une recherche *appliquée* (recherche "finalisée"), visant à la prévision ou au conseil : déterminer à quel type de décideurs elle est destinée (qui sont les commanditaires de la recherche), quels sont leurs objectifs, quels sont les facteurs sur lesquels ils peuvent agir (action commerciale, décision administrative ou législative, information du public, publicité, etc.) ou dont ils doivent tenir compte dans leur action (facteurs socioéconomiques, par exemple), et quelle influence ce contexte a pu avoir sur les conditions et le déroulement de la recherche.

Règle 2 : *Définir la population étudiée.*

Cette indication figure en principe dans le texte du commentaire, ou dans le titre même du tableau. On commencera par vérifier s'il s'agit d'un échantillon, ou de l'ensemble de la population étudiée (cas des statistiques administratives, et des recensements). S'il s'agit d'un *échantillon*, on vérifiera la manière dont il a été constitué (sondage aléatoire ou par quotas, ou autre procédure), sa représentativité par rapport à la population dont il est extrait, et les risques d'erreur que l'on court en généralisant la description de l'échantillon à la population dans son ensemble (par exemple, limites de confiance, lorsque celles-ci sont calculables). Dans tous les cas : on vérifiera l'effectif de la population, on décrira ses principales caractéristiques (si l'unité statistique est l'individu, ce qui est de loin le cas le plus courant : âge, sexe, PCS, habitat, etc.) ; on s'assurera enfin qu'il n'y a pas eu sélection d'une sous-population ("filtrage" des seuls hommes chefs de ménage, élimination des "non répondants", etc.).

Règle 3 : *Ne pas confondre les dimensions sociologiques et leurs indicateurs.*

Il s'agit là d'une erreur courante, qui consiste à substituer au nom des variables l'interprétation sociologique que l'on compte en donner. C'est le cas lorsque, par exemple, on parle de "classes sociales" en décrivant un tableau utilisant la variable "catégories socio-professionnelles" définie par l'INSEE. Il est certes possible d'interpréter cette variable empirique en termes de strates, ou de classes (au prix de regroupements parfois délicats) ; mais il est prudent de bien distinguer la description des données, aussi "objective" que possible, avec l'interprétation (subjective) que l'on en propose.

Règle 4 : *Tenir compte des propriétés formelles des variables.*

Dans les regroupements, les représentations graphiques, et les calculs (pourcentages, valeurs moyennes, coefficients de liaison, etc.), prendre en compte le niveau de mesure des variables, et ne pas utiliser des procédures correspondant à un niveau supérieur. Il arrive que l'on s'autorise à enfreindre cette règle, et que par exemple l'on conçoive un "indice empirique" [1] faute d'un coefficient mieux adapté ; dans le traitement des variables ordinales, c'est le cas des "rangs moyens" ou des scores d'attitude moyens. Cette pratique, si elle n'est pas légitimée par les propriétés que l'on attribue aux variables, ne doit être utilisée qu'en première approximation.

Règle 5 : *Dans l'analyse, toujours aller du simple au complexe.*

Au niveau du contenu, analyser les marges avant de pénétrer plus avant dans le tableau. Ne pas hésiter à simplifier le tableau en procédant à des regroupements, à condition naturellement qu'ils soient pertinents (qu'ils aient un sens par rapport aux données, et aux questions que l'on se pose). Au niveau des procédures, commencer par utiliser les techniques les plus simples (calcul de pourcentages) ; s'il subsiste des problèmes non résolus, ne passer à des techniques plus raffinées qu'après avoir épuisé toutes les possibilités des techniques simples.

Règle 6 : *Dans l'analyse, commencer par les faits les plus saillants.*

Même si de légères différences ou de légères variations sont plus riches d'enseignement que les effets importants (qui peuvent être déjà bien connus des spécialistes), commencer néanmoins par ces derniers : ils permettent de replacer les faits intéressants dans leur contexte, et d'en donner la mesure exacte.

---

[1] Ou "indice énumératif", selon la terminologie de Raymond Boudon et Paul F. Lazarsfeld, *Le vocabulaire des sciences sociales*, Paris, Mouton, 1965, Pages 14-16, et 95-131.

Règle 7 : *Matérialiser les hypothèses explicatives par des modèles concrets.*

Traduire les hypothèses par un tableau du même type que le tableau analysé, en conservant les marges de ce dernier. Chaque tableau "théorique" (à marges constantes) constitue un *étalon*, par rapport auquel on situe les données d'observation. En l'absence d'autres éléments, le tableau "théorique" le plus proche du tableau analysé correspond à l'hypothèse la plus plausible. En outre, la confrontation des données d'observations à un tableau "théorique" construit à partir d'une hypothèse permet une analyse fine du tableau, procédant case par case.

Règle 8 : *Ne comparer que ce qui est comparable.*

Pour la comparaison de tableaux tirés d'enquêtes différentes, on ne retiendra que ceux pour lesquels les populations et les variables étudiées sont de même nature. S'il y a lieu, on procédera aux regroupements nécessaires afin de rendre les tableaux formellement identiques. Dans l'interprétation, on gardera cependant à l'esprit les inévitables disparités dues aux différences de culture ou d'époque.

Règle 9 : *Ne tenir compte que des indices que l'on sait interpréter.*

La littérature sociologique abonde en indices empiriques, conçus pour résoudre un problème spécifique. Avant de tirer des conclusions à partir des valeurs prises par les indices *ad hoc* utilisés par les auteurs, s'assurer que l'on en a bien compris la construction, les possibilités, et les insuffisances. S'il s'agit d'indices statistiques reconnus, mais peu familiers, ne pas hésiter à prendre l'avis d'un statisticien [1]. En cas de doute sur la signification d'un indice, il est prudent de ne pas tenir compte des informations qu'il est censé apporter.

Règle 10 : *N'utiliser soi-même que des indices éprouvés.*

On donnera la préférence aux indices statistiques usuels (en n'utilisant que ceux dont on a bien assimilé les propriétés). Si l'on décide de recourir à un indice *ad hoc*, on veillera à ce qu'il ne mette en jeu que des opérations dont les implications sociologiques sont évidentes. Dans tous les cas, on s'assurera que l'indice utilisé n'est pas en contradiction avec les théories sociologiques auxquelles on se réfère ; par exemple, si l'on estime que la société que l'on décrit est une société conflictuelle et fortement hétérogène, calculer un revenu moyen est certes possible, mais c'est un contresens sociologique.

[1] Le dialogue entre sociologues et statisticiens n'est pas toujours exempt de malentendus. L'avis sollicité doit porter sur les propriétés *formelles* de l'indice : la manière dont il a été conçu, ses limites de variation, la signification des valeurs remarquables qu'il peut prendre (minimum, maximum, zéro, valeurs négatives, valeur infinie, etc.). Par contre, les implications sociologiques des valeurs prises par l'indice sont naturellement de la compétence du seul sociologue.

Règle 11 : *Ne pas confondre caractéristiques individuelles et caractéristiques collectives.*

En sociologie électorale, on ne dispose pas d'informations sur le scrutin au niveau individuel, mais seulement au niveau des circonscriptions [1]. Pour étudier les relations entre la PCS et le comportement électoral, on met en parallèle les résultats du scrutin et la composition sociale de chaque circonscription ; cela permet de dire que telle PCS vote préférentiellement pour tel courant politique. Dans la plupart des cas, cette conclusion est valide, car confirmée par d'autres informations, portant sur les individus (résultats de sondages ou d'enquêtes qualitatives). Mais en l'absence d'autres sources d'information, de telles corrélations sont à manier avec prudence : si l'on constate une corrélation positive entre la proportion d'immigrés par circonscription et le vote d'extrême droite, doit-on en conclure que les immigrés votent préférentiellement pour ce courant politique ?

Règle 12 : *Se méfier des interprétations abusives.*

Bien distinguer le niveau descriptif, qui peut reprendre les catégories et les termes utilisés par les auteurs du tableau, et les niveaux interprétatifs, qui renvoient à des théories sociologiques. Ne pas confondre par exemple l'intention de vote déclarée à un enquêteur, le vote lui-même, et les déclarations sur ce vote faites à un enquêteur après la publication des résultats du scrutin. S'efforcer de bien comprendre le sens que l'auteur donne aux mots qu'il utilise, et de bien voir en quoi ce sens diffère de la signification qu'on lui attribue soi-même. Bien distinguer la notion abstraite, qui fait référence à une théorie sociologique (par exemple : la mobilité sociale), et le phénomène concret que l'on considère comme son indicateur (par exemple : la mobilité professionnelle intergénérationnelle).

Règle 13 : *Interroger les données, en leur posant des questions précises et circonstanciées.*

C'est la règle fondamentale qui préside à toute analyse de données ; cela signifie que, devant d'un tableau statistique, on ne peut se contenter de viser à une interprétation globale et exhaustive par le seul survol du tableau (ce qui serait complètement utopique) ; il faut adopter un comportement *actif* de recherche des réponses aux questions que l'on se pose. Pour assimiler les informations contenues dans un tableau, il faut d'abord les déconstruire, en identifiant la problématique de l'auteur du tableau, puis les restructurer en fonction de sa problématique personnelle. Après avoir pris connaissance des caractéristiques d'un tableau, on voit assez bien quel type d'informations il peut apporter, quelles sont, à leur propos, les questions intéressantes, et comment obliger le tableau à livrer ces informations (dans le respect des propriétés des variables).

[1] Les "sondages à la sortie des urnes" (SSU) compensent partiellement cette impossibilité, mais avec deux limitations : il s'agit d'enquêtes sur échantillon (donc non exhaustives), et les informations recueillies sont de simples déclarations des votants (donc moins fiables que les votes eux-mêmes).

b) Lectures complémentaires.

Pour en savoir plus sur la lecture des tableaux statistiques, on se reportera à : Michel Novi, *Pourcentages et tableaux statistiques* (Paris, PUF, 1998). On trouvera dans ce *Que sais-je ?* des précisions et de nombreux compléments techniques qu'il n'était pas possible d'inclure dans ce manuel.

Sur la mise en œuvre de modèles explicatifs à confronter aux données d'observation, on pourra lire : Raymond Boudon, *Les mathématiques en sociologie* (Paris PUF, 1971). Ce petit ouvrage présente l'analyse causale (*path analysis*), méthode de description des liaisons entre variables à partir d'hypothèses sur les relations de causalité entre ces variables.

L'utilisation des indices, leurs propriétés, et leur rapport avec les théories sociologiques font l'objet d'un choix de textes d'auteurs variés, sélectionnés et présentés par Raymond Boudon et Paul F. Lazarsfeld : *Le vocabulaire des sciences sociales. Concepts et indices* (Paris, Mouton et MSH, 1965). On pourra compléter par, des mêmes auteurs : *L'analyse empirique de la causalité* (Paris, Mouton et MSH, 1966).

Un certain nombre d'usages abusifs ou fallacieux des statistiques, tirés pour la plupart d'articles de presse, sont recensés dans : Joseph Klatzmann, *Attention, statistiques ! Comment en déjouer les pièges* (Paris, La Découverte, 1992), ainsi que dans : Association Pénombre, *Chiffres en folie* (Paris, La Découverte, 1999).

Enfin, on ne saurait trop conseiller au lecteur de se reporter d'une part aux livres consacrés aux mathématiques et aux statistiques appliquées aux sciences de l'homme, et d'autre part aux ouvrages d'épistémologie et de méthodes des sciences sociales. Ils sont trop nombreux pour pouvoir être énumérés ici.