UNIVERSITE de ROUEN
S.C.U.R.I.F.F.

I R E M
de
ROUEN

1, rue Thomas Becket - BP 153
76135 Mont-Saint-Aignan Cédex
tél : 35 14 61 41

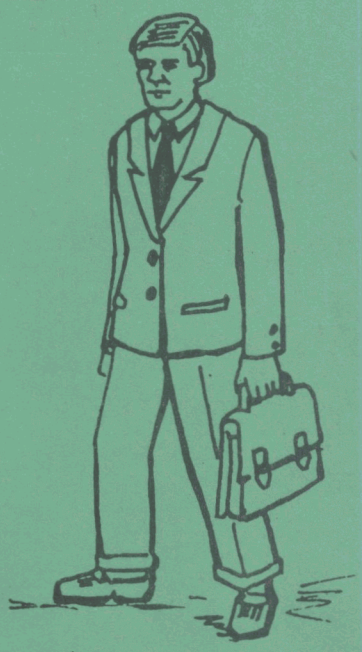
Dossier réalisé par:
B. LANNUZEL
G. ORANGE
J. F. PICHARD
octobre 1989

LES ENQUETES

A QUESTIONS NOMINALES

homme du peuple
1789

Réflexions
et
méthodologie
pour l'exploitation
d'une enquête
à questions nominales



homme moyen
1989

UNIVERSITE de ROUEN
S.C.U.R.I.F.F.

I R E M

de

ROUEN

1, rue Thomas Becket - BP 153
76135 Mont-Saint-Aignan Cédex
tél : 35 14 61 41

Dossier réalisé par:

B. LANNUZEL
G. ORANGE
J. F. PICHARD

octobre 1989

LES ENQUETES

A QUESTIONS NOMINALES

homme du peuple
1789

Réflexions
et
méthodologie
pour l'exploitation
d'une enquête
à questions nominales

homme moyen
1989

Les Français sont comme ça !

Un sondage sur les 16-24 ans

Jeunes, individualistes, généreux...

SE MARIER UN JOUR

● Croyez-vous qu'un jour, vous vous marierez ?

● Pensez-vous que vous aurez un jour des enfants ?

VIVRE EN MUSIQUE

● La musique occupe-t-elle dans votre vie une place :

Le Français moyen n'existe pas

LA CROIX 3/2/82

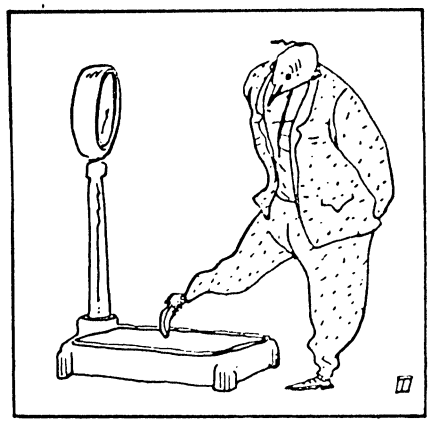
Les Français disent ce qu'ils pensent de la chasse

LE PRIX DE VOTRE ENFANT

Selon une étude de l'INSEE, un enfant coûte en moyenne 1760 francs par mois. Mais la charge varie selon l'âge, le nombre de frères et de sœurs, et les ressources des parents (P. 12 et 13)

Le coût de la rentrée scolaire

Il faut compter, pour les familles, selon l'INSEE, 1.760 F en moyenne par enfant chaque mois, habillement et loisirs compris.



Le budget des familles au crible de la statistique

Le troisième enfant coûte plus cher

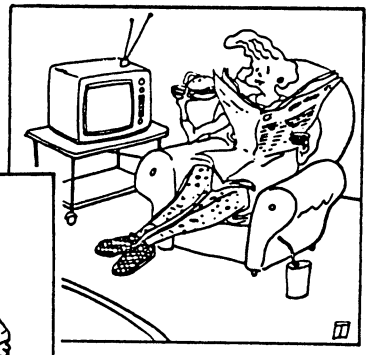
Variations sur un budget

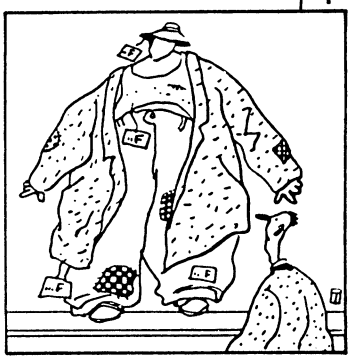
Les dépenses pour l'alimentation prennent plus d'importance dans les ménages à faibles ressources ; chez les cadres ce sont les vacances... selon une enquête de l'INSEE

Les gros portefeuilles travaillent moins...

Selon un sondage SOFRES- « *Marie-Claire* »

L'argent ne fait (toujours) pas le bonheur



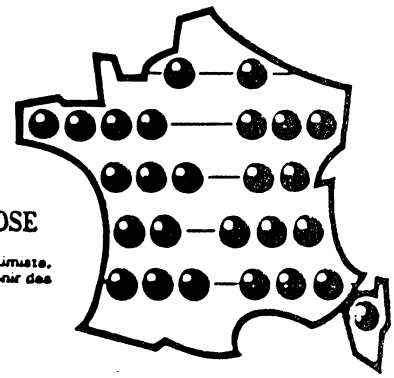


Nos femmes portent la culotte

Consommation : la voiture au premier rang

LE MATIN 3/2/1982

une belle moisson de chiffres pour vous aujourd'hui... et vos enfants demain!



L'AVENIR EN ROSE

● Etes-vous plutôt optimiste, plutôt pessimiste sur l'avenir des jeunes en France ?

LA PEUR DE LA GUERRE

● Pourriez-vous dire ce qui vous fait le plus peur dans l'avenir ?

31e RECENSEMENT DE LA POPULATION

SOMMAIRE

PRESENTATION

Ce fascicule doit son origine à une constatation faite par tout le monde : la place importante, quelquefois trouvée envahissante, des statistiques dans la vie quotidienne !

Enquêtes d'opinion à domicile, sondages pré-électoraux, constitution de panels pour estimer l'audience des chaînes de télévision, diffusion de questionnaires dans les établissements scolaires, voici quelques-unes des pratiques auxquelles nous sommes couramment confrontés, en dehors des multiples études économiques, démographiques et sociales.

Parfois jugées comme des agressions, parfois comme des moyens de pression ou de manipulation, les statistiques sont le plus souvent acceptées comme inéluctables.

Les instituts de sondage (IPSOS, SOFRES,...) sont généralement les maîtres d'oeuvre de telles enquêtes. Considérés comme des instituts sérieux utilisant des méthodes scientifiques, ils sont de ce fait peu contestés. Mais les études réalisées par ces instituts sont, pour la plupart d'entre elles, demandées par des groupes de presse ou commerciaux et les médias (journaux, radios et télévisions) sont les intermédiaires obligés entre les statisticiens et le public. C'est donc souvent aux journalistes qu'appartient la tâche de diffuser les résultats des enquêtes, de juger les questions intéressantes pour leurs lecteurs, et de commenter les dits résultats.

Cependant, les instituts spécialisés ne sont pas les seuls à mettre en oeuvre des investigations statistiques: les chercheurs en Sciences Humaines, les enseignants, les responsables de collectivités locales, etc..., ont recours eux aussi à ces méthodes dans l'espoir de récolter des informations utiles pour leur connaissance des élèves, du corps électoral... ou pour des prises de décision comme la création d'un centre commercial ou culturel, l'itinéraire d'un ramassage scolaire...

Il est possible, soit en tant que "consommateur", soit en tant que "réalisateur", que l'on se sente démuni devant de telles statistiques et que l'on désire avoir une approche des techniques utilisées.

Ce document a donc l'ambition de donner quelques éléments de théorie, de réflexion et de pratique pour mieux comprendre certains problèmes statistiques relatifs essentiellement aux enquêtes à questions nominales.

Nous envisagerons en particulier les deux démarches susceptibles d'être suivies par le statisticien selon que sa préoccupation est une meilleure connaissance de la population (par l'intermédiaire des caractères retenus) ou l'étude des relations entre les variables (par l'intermédiaire des observations faites sur certains individus).

Nous examinerons aussi l'attitude, entretenue par les médias, qui consiste à utiliser les sondages pour tenter de dresser le portrait d'un individu-type sensé représenter la population toute entière. Cette démarche, qui aboutit à définir un "homme moyen", gomme les disparités souvent significatives et soulève de ce fait un double problème, celui de la pertinence de cette démarche et celui du choix des méthodes utilisées pour parvenir à une telle représentation.

Par contre, nous n'aborderons pas ici les travaux qui précèdent les traitements, à savoir la rédaction du questionnaire (choix et formulation des questions), le choix et la qualité des échantillons éventuels (représentativité), et les tâches relatives au recueil et à la saisie des informations.

Nous aimerions que cette brochure soit un document de travail et nous serions heureux de recevoir commentaires, critiques, idées et suggestions pour la poursuite et l'amélioration de ce texte.

PREAMBULE

UN PEU D'HISTOIRE

Dès le 5e millénaire avant J.C., des dénombrements sur de vastes populations sont réalisés dans les empires de l'Antiquité qui se constituaient. Les données sont très fragmentaires, mais les documents parvenus à ce jour permettent d'affirmer qu'en Egypte, vers 3000 avant J.C., il y eut un recensement des hommes pour satisfaire les besoins en main-d'oeuvre dûs à la construction des pyramides, et un recensement des biens et richesses pour établir des impôts ; d'autres eurent lieu à la même époque dans l'empire de Sumer et en Mésopotamie.

Dans l'Empire du Milieu, en Chine vers 2200 avant J.C., on trouve un inventaire des hommes et des terres, après une grande inondation, pour répartir les terres. Chez les Hébreux, vers 1000 avant J.C., un dénombrement est ordonné par Moïse après la sortie d'Egypte (Bible, Nombres I).

A partir du 11e siècle avant J.C., des recensements réguliers sont effectués en Chine, en Perse, en Egypte pour établir les rôles d'impôts, de corvées ou de conscription militaire. Il en sera de même dans l'empire romain au 6e siècle avant J.C., en Inde où un ministre écrit au 4e siècle avant J.C. dans un traité : "*l'état doit tout diriger et contrôler, et pour cela connaître par recensement la population, animaux, matières premières, prix et salaires, ...*", au Japon, dans l'empire Maya, etc...

Ainsi les grands empires de l'Antiquité, centralisateurs et unificateurs, ont ressenti très tôt la nécessité de dénombrer les hommes et leurs biens pour administrer au mieux leurs territoires et répondre aux besoins humains, matériels et financiers créés par les guerres et les travaux publics.

L'effondrement de l'empire romain marque la fin des recensements généraux et périodiques en Europe. Du 5e au 16e siècle, seules des opérations partielles et occasionnelles (capitulaires) sont effectuées :

- le "Domesday book" en Angleterre en 1086, puis en 1348 après la grande épidémie de peste noire.

- l'Etat des paroisses... de 1328 en France et d'autres relevés pour établir l'assiette des impôts, lever des corvées ou des armées, ce qui explique l'hostilité des habitants.

Les enregistrements d'état-civil (baptêmes, mariages, décès) commencent à la fin du 14e siècle, mais les registres furent très mal tenus jusqu'au milieu du 18e siècle.

Au 16e siècle, à l'imitation de Platon, des ouvrages de géographie et de sciences politiques sont publiés, et certains auteurs, tel J. Bodin dans son livre La République en 1576 demande la "*censure qui est l'estimation des biens de chacun*" et poursuit "*quant on saura le nombre, l'âge et la qualité des personnes, on saura combien on pourra en tirer pour la guerre, les colonies ou les travaux publics. Le dénombrement des biens est indispensable afin qu'on sache les charges que chacun doit porter.*"

Au 17e siècle, le besoin de connaître et d'expliquer les phénomènes économiques et sociaux fait apparaître les recensements non seulement comme un instrument de connaissance pour une action immédiate telle la levée des impôts ou des armées, mais aussi de compréhension pour mieux gouverner. L'idée d'*arithmétique politique* c'est-à-dire la modélisation mathématique de ces phénomènes est lancée à la fin du 17e siècle en Angleterre en particulier par Graunt et Petty, et exposée en 1705 par Jacques Bernoulli dans son livre Ars Conjectandi. Elle va se propager au 18e siècle dans tous les pays européens avec Leibniz en Allemagne, Moivre en Angleterre, Buffon, Condorcet et Hancarville en France, Kersseboom en Hollande, Plà en Italie....

Dans l'Encyclopédie de Diderot et d'Alembert, l'arithmétique politique est décrite comme "*l'art de réduire aux principes du calcul les principaux objets de gouvernement de l'Etat*", et pour cela nécessite d'avoir une description détaillée des hommes et des biens. Cependant, à cause de l'hostilité des populations et après plusieurs tentatives aux 17e et 18e siècles, des recensements généraux et réguliers ne furent effectués qu'en 1749 en Suède, en 1797 en Espagne, en 1800 en Angleterre, en 1801 en France...

Prenant appui sur les méthodes mathématiques et probabilistes des moindres carrés et de la loi normale, étudiées au début du 19e siècle par Laplace, Gauss et Poisson, et prolongeant l'anthropométrie fondée en 1835 par l'astronome et statisticien belge Quételet, l'école anglaise de biométrie et de psychométrie, avec Galton, K. Pearson, R. Fisher..., va développer de nouveaux outils statistiques (corrélation, analyse factorielle, χ^2...).

Les statistiques sont maintenant utilisées dans tous les domaines, démographique, économique et social en particulier, pour permettre une meilleure compréhension de nos sociétés afin de prévoir leurs évolutions ou même d'avoir une influence sur celles-ci.

Cette brève évocation montre qu'on est ainsi passé du "connaître pour agir" des empires et royaumes, de l'Antiquité jusqu'au 18e siècle, au "connaître pour comprendre et agir mieux", objectif actuel qui passe par une modélisation, un recueil et un traitement pertinents des observations en fonction du but poursuivi.

On pourra consulter entre autres: Benzécri J.P., Histoire et préhistoire de l'analyse des données, Dunod, Paris, 1982.

CHAPITRE 1

CONNAITRE UNE POPULATION

1.1. Population et Individus

Toute étude statistique nécessite que l'on définisse très précisément la population sur laquelle elle porte. Par *population* nous entendrons, dans ce document, un ensemble fini, noté P, dont les éléments -objets ou personnes-, appelés *individus* ou *unités statistiques*, possèdent un certain nombre de caractéristiques que l'on désire étudier.

Le nombre des individus, ou *effectif* de P, est noté N et nous supposons que tous les individus sont munis d'un "poids" égal (situation d'*équipondération*) et qu'aucune structure mathématique n'est définie sur P. En particulier on ne retient pas la relation d'ordre liée au temps si les unités statistiques sont des instants -jours, années...-, ni la relation spatiale si les individus sont localisés dans l'espace -régions, parcelles de terrain...-

1.2. Recensement ou sondage

Pour collecter des informations sur une population, on a recours à des enquêtes basées sur des questionnaires, à des observations physiques, à des expériences dont on relève les résultats, à des bulletins de vote, etc...

Une telle collecte peut être *exhaustive* si elle porte sur tous les éléments de la population. C'est le cas des recensements :

- Recensement national (en 1982: 31 ^{ème} recensement général en France).
- Référendum (recensement selon un seul caractère)
- Election présidentielle en France (en 1988: choix parmi 14 candidats...)
- Statistiques rectorales :
    - Recensement des élèves de l'Académie
    - Recensement des élèves présentés au baccalauréat,
    - ...
- Relevés géographiques ou économiques: taux de chômage par département....

Remarquons toutefois qu'il est souvent difficile d'atteindre tous les individus sans exception à cause généralement de difficultés matérielles : absence de réponse ou de vote, individus échappant au recensement... Si la proportion des individus "absents" est faible et ne perturbe pas trop les jugements sur la population, on qualifiera la collecte de *quasi-exhaustive*.

La collecte peut être *partielle* si elle ne porte que sur une partie des individus : c'est le cas en particulier des enquêtes d'opinion par sondage, ou du contrôle de qualité sur un lot de fabrication. La sous-population des individus choisis pour l'étude est appelée dans ce cas un *échantillon*.

Nous étudierons ici quelques méthodes d'exploitation des résultats recueillis sur un ensemble d'individus sans désir d'étendre les jugements portés à d'autres individus qu'à ceux qui ont été questionnés ou observés. Le problème de l'échantillonnage et/ou de la représentativité dans le cas d'une collecte partielle n'y sera donc pas soulevé.

Nous utiliserons -à titre d'exemples- soit des données d'enquêtes exhaustives, soit des résultats de sondages sur échantillon parus dans la presse. Dans ce dernier cas, nous ferons donc abstraction -sauf mention rapide- des problèmes de représentativité qui conduisent à nuancer les conclusions projectives ou inférentielles faites sur la "population-mère".

1.3. Caractères et modalités

La connaissance de la population passe par le relevé de certaines caractéristiques.

Pour relever une caractéristique, il faut au préalable préciser les différentes valeurs qu'elle peut prendre, compte tenu en particulier des "instruments de mesure" dont on dispose: questions dans une enquête, appareil physique..., pour saisir l'information. Les valeurs "observables" constituent l'ensemble des *modalités*.

Si M est un tel ensemble de modalités, on appelle *caractère* C défini sur P et à valeurs dans M, toute application qui à chaque individu associe une et une seule modalité.

Selon la nature de l'ensemble des modalités -ensemble structuré ou non, ensemble numérique ou non...- divers types de caractères peuvent être considérés: nominaux, ordinaux, numériques. Nous nous intéresserons, dans les chapitres suivants, aux caractères dits "*nominaux*".

1.4. Nature des caractères

Dans une enquête, toutes les questions ne sont pas de même nature :

- certaines sont destinées à comprendre la composition ou la structure de la population, indépendamment du but de l'enquête ;
- d'autres sont relatives à l'enquête proprement dite.

Considérons une enquête d'opinion dans laquelle on demande aux personnes interrogées leur sexe et leur âge. Il va de soi que le sexe ne peut expliquer l'âge et vice versa.

On ne peut mettre sur le même plan des réponses dont les unes portent sur le sexe ou l'âge et les autres sur la question: "*pensez-vous que les riches sont plus heureux* ?". En d'autres termes, le sexe ou l'âge doivent éclairer l'opinion prise et non l'inverse.

Les caractères "*structurels*" qui ne varient pas avec le temps ou varient de façon prévisible, tel l'âge, s'opposent aux caractères "*conjoncturels*" dont l'évolution dépendra de mille facteurs et qui traduisent un état d'esprit du moment et non pas un état peu ou prou immuable.

Bien sûr, cette distinction est relative à une étude donnée et ne repose sur aucune propriété mathématique des caractères.

Dans la pratique, le partage n'est pas toujours aisé et certaines questions se situeront à la frange. Le choix est fonction du but recherché et devient du ressort du praticien.

1.5. La finalité d'un recensement

Certains recensements sont conçus en fonction d'un but précis: un référendum donnera la réponse majoritaire à une question, une élection désignera -en fonction des procédures prévues- le ou les candidats élus...

Autre cas simple, celui où l'étude porte sur quelques critères structurels: sexe, âge, catégorie socio-professionnelle..., sans volonté d'explication mais dans le souci de connaître une stratification (1) de la population. Le but peut être alors la construction d'échantillons représentatifs en vue de sondages par quotas.

---

(1) stratification: décomposition d'une population en sous-populations homogènes par rapport à une caractéristique.

Comme toute activité humaine, la statistique n'échappe pas au principe de finalité: l'adaptation des moyens à des fins. Les premiers recensements de population avaient un but politique. -précisément fiscal ou militaire...

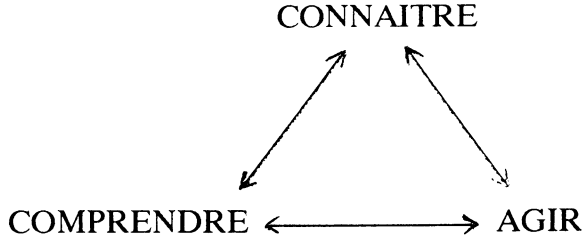
De nos jours, où les recensements sont légion, nous devons nous interroger plus en détail sur cette finalité. A l'instar de toute science, nous pouvons assigner à la statistique un but oscillant entre *connaissance* et *action*. Encore que l'on puisse passer de l'une à l'autre selon deux démarches différentes :

- soit la simple description des données collectées débouche aussitôt sur une décision d'action, c'est la démarche politique;

- soit ces données nécessitent une transcription selon un modèle préalable conduisant, au-delà de la simple connaissance des phénomènes, vers une compréhension plus profonde, c'est la démarche scientifique.

Cette dernière démarche peut conduire à l'action mais aussi trouver là sa finalité. La statistique historique est bien source de description et de compréhension de phénomènes anciens mais ne s'inscrit dans aucune praxis.

Le schéma suivant illustrera notre propos :



La mise en évidence de ces trois pôles interdépendants peut seule éclairer le choix des techniques utilisées, celles-ci étant assujetties au but poursuivi.

La primauté du conceptuel sur le calcul est absolue malgré les tendances actuelles qui grâce au recours à l'informatique relèguent les aspects essentiels des choix de traitement statistique au simple maniement de fonctions que l'ordinateur appliquera bien sûr sans réflexion !

Le premier pôle -étape obligée- consiste à CONNAITRE.

A ce stade, la finalité est simplement la description de la population suivant certains caractères préalablement choisis. On dénombrera -par exemple- les "demandeurs d'emploi" selon le sexe, l'âge... Les techniques utilisées -tels le *tri-à-plat*, le *tri-croisé*, et des représentations graphiques judicieuses (2)- apporteront une connaissance purement descriptive et rien de plus.

A un deuxième niveau ou pôle, le statisticien cherchera à COMPRENDRE. Les questions peuvent être les suivantes:

- existe-t-il des relations entre les caractères ?

- l'observation de la modalité d'un des caractères (femme) modifie-t-elle la répartition d'un autre caractère (la durée du chômage) ?

- peut-on dégager dans la population des familles d'individus semblables, au comportement voisin ?

Ces questions ne peuvent trouver de réponses qu'au moyen d'un jeu d'hypothèses constituant un modèle, dont le choix subjectif voire arbitraire, sera suggéré généralement par la description faite antérieurement. Le modèle -qui utilise le plus souvent des hypothèses de linéarité ou de structures euclidiennes, ou des hypothèses probabilistes ("normalité" des lois)...- résultera de l'examen approfondi de la description des données, mais aussi de la réflexion, de l'intuition ou même parfois de l'utilisation du logiciel disponible.

Le dernier pôle échappe en principe au statisticien mais relève du politique: AGIR... sur les individus car il est impossible d'agir sur les caractères : sexe, âge... ou opinion! Sur ce dernier point la confusion est courante : on ne peut agir sur l'opinion pour qu'elle change, mais on agit éventuellement sur l'individu pour qu'il change d'opinion !

La nuance est de taille !

1.6. Pertinence des données

Les prévisions d'actions et l'expérience accumulée antérieurement détermineront le choix des caractères étudiés, de certains traitements, de certaines descriptions, même si ces choix ne sont pas "neutres".

---

(2) en particulier les diagrammes

- en barres (parfois appelé improprement histogramme)
- en secteurs (appelé aussi en fromage ou camembert)

Certaines règles doivent néanmoins respecter un souci de complète information de nature à éviter des erreurs d'interprétation grossières, suggérées par une information vraie mais partielle, ou subtiles. La qualité d'une exploitation statistique va dépendre de la qualité du matériau brut : l'information élémentaire. Certains principes doivent être retenus :

- délimitation du champ d'étude, c'est-à-dire des contours exacts de la population en déterminant quels individus en font partie à l'exclusion de tous les autres individus possibles. La fiabilité de l'étude, pour l'action qui en résultera, dépendra bien sûr de la pertinence et de l'homogénéité de ce champ d'étude.
- choix des caractères et de leurs modalités qui donneront forme à des questions précises et adaptées, évitant si possible tout biais en assurant la neutralité de la formulation.
- sérieux du recueil, c'est-à-dire du relevé sur le terrain des réponses de chaque individu.
- soin apporté au dépouillement.

Nous n'aborderons que fort peu ces aspects préparatoires pour nous en tenir au traitement, qui relève plus précisément du statisticien. Mais ces étapes préalables déterminent la confiance que l'on accordera aux résultats obtenus.

Précisons que la taille de la population -cent, mille ou un million d'individus- change peu la complexité de l'étude. Seuls les travaux préparatoires de recueil et de dépouillement sont alourdis. Le statisticien maniera alors des nombres plus importants ! En revanche, le nombre des caractères soumis à son examen et surtout le nombre de modalités et leur nature -numérique ou nominale- peut compliquer son travail au point de le rendre impossible. C'est pourquoi nous insisterons sur le traitement des caractères dichotomiques : oui/non, vrai/faux, d'accord/pas d'accord...

Revenons plus en détail sur notre deuxième niveau de finalité : comprendre les lois régissant les caractères ou le comportement des individus.

La recherche peut tout aussi bien s'appuyer sur les individus pour appréhender les liens entre les caractères ou au contraire partir des caractères pour étudier les relations entre les individus. La finalité impose l'une ou l'autre des approches, parfois les deux !

Un exemple mettra cette particularité en évidence : un botaniste a relevé, sur plusieurs territoires, les espèces végétales présentes et divers autres caractères. La finalité de l'étude pourra être soit l'étude des caractéristiques des espèces relativement à leur milieu naturel -le territoire-, soit une typologie des territoires au regard des espèces présentes. Du point de vue mathématique, le passage d'une optique à l'autre est toujours possible, mais pas toujours pertinent. Tout dépendra de l'orientation que l'on veut donner à la recherche.

CHAPITRE 2

CARACTERES NOMINAUX

2.1. Définitions

Un caractère défini sur une population P est **nominal** lorsque l'ensemble de ses modalités (l'ensemble des réponses possibles à une question, ou l'ensemble des valeurs possibles pour une caractéristique) est fini et n'est muni d'aucune relation d'ordre naturelle ; c'est-à-dire les réponses sont distinctes mais on ne peut pas dire que l'une est "plus grande", ou "meilleure" qu'une autre ou "avant" une autre.

Exemples :

1. On demande à une population de bacheliers la série de leur bac : A,B,C,D,D',E,F

2. On demande à des électeurs le candidat de leur choix (en précisant la liste des candidats)

...

Lorsque l'information est obtenue lors d'une enquête par questionnaire, un caractère nominal correspond à une question fermée à choix unique dont la liste des réponses possibles est soit imposée par l'enquêteur, soit fixée par la nature du caractère. Une telle question sera dite aussi nominale.

Sont bien entendu exclues les questions formulées en des termes tels que : "*cochez une ou plusieurs cases selon que vous êtes d'accord ou non avec les différentes opinions suivantes...*" (questions à choix multiples) ainsi que les questions conditionnelles telles que : "*Si vous avez répondu oui à la question précédente, que pensez vous...*"

Caractères nominaux et caractères qualitatifs

Les caractères nominaux sont parfois appelés "qualitatifs", en ce sens qu'ils s'opposent aux caractères "quantitatifs" (numériques).

Des différences peuvent toutefois apparaître entre les définitions :

- le terme qualitatif peut être attribué à des caractères dont la seule caractéristique est de ne pas avoir des modalités numériques ; on peut ranger sous ce terme des caractères "ordinaux" (munis d'un ordre naturel entre les diverses modalités) tels que: "*Avec l'opinion suivante..., êtes-vous: pas du tout, un peu, ou tout à fait d'accord ?*"

- ce qui importe n'est pas la nature des relevés : le fait d'utiliser des "chiffres" pour repérer les réponses d'une question ne conduit pas nécessairement à classer le caractère comme "quantitatif" si ces chiffres sont utilisés, non comme des nombres, mais comme des codes ou des symboles.

Caractères dichotomiques et polytomiques

Un cas particulier important et fréquent est celui où le caractère n'a que deux modalités. On dit alors qu'il est *dichotomique* ou *booléen*. Certains statisticiens appellent parfois un tel caractère un attribut.

C'est le cas par exemple :

- du caractère "sexe"
- des questions où les réponses possibles sont "oui/non", ou "d'accord/pas d'accord"...
- des caractéristiques "présence/absence" (par exemple d'une espèce végétale), "bon/mauvais" (pour des pièces d'un lot de fabrication)...

Si un caractère nominal a plus de 2 modalités, il est dit *polytomique*.

2.2. Effectifs et proportions

Intéressons-nous à une population P formée de N individus et à un caractère nominal C, défini sur cette population, prenant les k modalités $(c_1,c_2,..,c_k)$.

L'ensemble des individus ayant la modalité c_i forme une *sous-population* notée C_i, et le nombre n_i positif ou nul de ces individus est l'effectif de la modalité c_i ou de la classe C_i. Le caractère nominal C induit une partition sur l'ensemble des individus dont les classes sont les sous-populations C_i.

Les effectifs vérifient donc : $\sum_i n_i = N$

L'importance de la classe C_i est mesurée par son effectif n_i mais dans certains problèmes (comparaison des classes, comparaison d'enquêtes...) il est utile de mesurer cette importance par un "effectif relatif", obtenu en divisant l'effectif de la classe par l'effectif total de la population.

On appelle *proportion* de la modalité c_i le nombre noté p_i compris entre 0 et 1, égal au rapport entre n_i et N : $p_i = n_i/N$.

Ce nombre p_i est encore appelé *fréquence* de la modalité c_i.

Il est usuel aussi de mesurer cette importance par cette valeur p_i multipliée par 100, c'est-à-dire par un nombre décimal appelé le *pourcentage* de la modalité c_i.

Toutefois, l'utilisation des proportions ou des pourcentages n'est pas sans danger car l'information contenue dans ces valeurs est moindre que celle fournie par les effectifs eux-mêmes.

Précisons cette différence dans l'exemple suivant:

Lors d'une enquête sur le tabagisme dans les lycées, après avoir précisé ce que l'on entend par "fumeur", on peut formuler l'observation de diverses manières. Par exemple:

"15 élèves de seconde parmi les 34 observés sont fumeurs"

Cette constatation se limite à décrire la réalité sans souci de comparaison ni d'extension.

"le rapport du nombre d'élèves fumeurs observés en seconde au nombre d'élèves de seconde est de 15 à 34"

Cette affirmation est encore vraie, mais laisse apparaître que la fraction 15/34 peut servir de caractéristique de la classe.

Comme une fraction évoque un nombre rationnel on dira:

"la proportion des élèves fumeurs en seconde est à peu près égale à 0,441 "

Donner ce nombre n'est pas, bien entendu, équivalent à donner la fraction 15/34 car il aurait pu être obtenu de bien d'autres façons: si on avait observé 30 fumeurs parmi 68, ou 150 parmi 340, etc...

Aussi se donne-t-on le droit d'envisager que ce nombre aurait pu être obtenu sur une observation de 100 élèves. Dans ce cas, le nombre de fumeurs aurait du être de 44,1 (le fait que ce nombre n'est pas entier est une convention facilement admise).

On affirme ainsi:

"le pourcentage observé de fumeurs en seconde est de 44,1%"

Cette affirmation peut laisser penser que l'information a été recueillie sur un groupe type de 100 individus et pourra entraîner des abus lors de comparaison de pourcentages sur des sous-populations d'effectifs différents.

2.3. Tri-à-plat et répartition

Il y a quelques années, les informations relatives à chaque individu étaient relevées soit sur des fiches, soit sur des cartes perforées. Actuellement, elles sont en général enregistrées dans des fichiers informatiques stockés sur des supports magnétiques telles les disquettes, les bandes,....

Quel que soit le mode d'enregistrement des informations, on appelle *tri-à-plat* des individus selon le caractère C l'opération qui consiste à lire l'ensemble des fiches ou l'ensemble des enregistrements du fichier, et à comptabiliser les effectifs relatifs aux diverses modalités du caractère C.

Le résultat de ce tri est un tableau donnant la correspondance entre les modalités et les effectifs de ces modalités ; c'est la *répartition* des effectifs du caractère C:

Modalités	c_1	c_2	...	c_k
Effectifs	n_1	n_2	...	n_k

où la somme des effectifs n_i représente le nombre total des individus interrogés.

La suite (c_i) des modalités étant fixée, on peut associer au caractère C la suite numérique de k entiers :

$$t = (n_1, n_2, ..., n_k) \text{ avec } \sum_i n_i = N$$

Une telle suite est appellée un type de dénombrement et parfois aussi une "**statistique**" des N individus en k classes.

Plutôt que de présenter un tableau d'effectifs, il est fréquent de présenter les résultats dans un tableau où les effectifs sont remplacés par des proportions ou par les pourcentages correspondants :

Modalités	c_1	c_2	...	c_k
Proportions	p_1	p_2	...	p_k

où p_i représente la proportion ou le pourcentage, parmi les N individus, de ceux qui ont choisi la modalité a_i, c'est-à-dire $p_i = n_i/N$.

Rappelons que fournir les proportions n'est pas équivalent à fournir les effectifs sauf si on adjoint aux proportions l'effectif total de la population qui permet de reconstituer les différents effectifs.

Les tableaux en pourcentages sont d'une meilleure lisibilité; par exemple une valeur supérieure à 50% est le signe d'un comportement majoritaire. Ils permettent aussi une comparaison plus facile entre deux populations. Notons qu'il est fait obligation aux instituts de sondages d'indiquer le nombre total de personnes interrogées (d'où l'encart qui accompagne les résultats d'enquête), mais que cette information n'est pas étendue aux différents quotas, ce qui ne permet pas en général de reconstituer tous les effectifs.

2.4. Etude simultanée de deux caractères dichotomiques

(situation 2x2)

a - En guise d'exemple

Une enquête, faite auprès d'un groupe de 100 jeunes, comporte, entre autres questions, les deux suivantes :

Question A : *"Du mariage, diriez-vous que c'est quelque chose de dépassé ?"*

Question B : *"Si cela s'avérait utile pour votre carrière professionnelle, accepteriez-vous de passer plusieurs années à l'étranger ?"*

Comme nous l'avons vu au chapitre précédent, ces deux questions peuvent être triées à-plat et les résultats publiés sous la forme suivante :

Question A : Oui : 32 Non : 68

Question B : Oui : 59 Non : 41

Si on se limite à ce traitement "caractère par caractère", le commentaire qui accompagnera ces résultats pourra être ainsi rédigé :

"Les jeunes sont majoritairement pour le mariage et acceptent l'idée de privilégier leur carrière professionnelle..."

Un tel commentaire laisse à penser que l'on a dégagé un portrait-type des jeunes, dans lequel une majorité d'entre eux devrait se reconnaître.

Mais ce faisant, on oublie de s'interroger sur le type de liaison entre les deux opinions. Pour rechercher cette liaison, on doit faire un *"tri croisé"*, qui pourrait conduire au tableau suivant:

		question B		
		oui	non	
question A	oui	29	3	32
	non	30	38	68
	Totaux	59	41	100

Ce tableau montrerait alors que le groupe de jeunes le plus important est constitué par les 38 qui ont donné une réponse négative aux deux questions posées. Ce qui donne une interprétation différente des résultats...

Dans ce cas, il se dégage 3 familles d'opinion d'importance comparable et l'affirmation citée plus haut n'est vraie que pour 30 parmi les 100 jeunes.

b - Formalisation de la situation 2x2

Considérons une enquête où N individus sont interrogés ou observés selon 2 caractères dichotomiques A et B.

Notons a_1 et a_2 les modalités du caractère A

b_1 et b_2 les modalités du caractère B

Les individus se répartissent donc globalement en 4 groupes ou classes.

Un tri croisé consiste à dénombrer les individus de l'échantillon qui appartiennent à chacune des 4 classes, donc à élaborer un tableau du type:

		question B		
		b_1	b_2	
question A	a_1	s	u	s+u
	a_2	v	t	v+t
	Totaux	s+v	u+t	N=u+v+s+t

où "s" est le nombre des individus qui ont répondu a_1 à la question A et b_1 à la question B, "u" le nombre des individus qui ont répondu a_1 et b_2 ....

On appelle "effectifs marginaux de A" les nombres des individus qui ont répondu a_1 et a_2 (sans référence à la question B), c'est-à-dire les nombres (s+u) et (v+t), résultats que l'on obtient par tri-à-plat sur la question A.

On définit de même les "effectifs marginaux de B", à savoir les nombres (s+v) et (u+t).

Cette situation dite "2x2" est un cas particulier de la situation qui consiste à étudier simultanément deux caractères nominaux.

2.5. Etude simultanée de deux caractères polytomiques:

Situation pxq

a - Exemple

[1] NOMBRE D'APPAREILS AUDIOVISUELS ET VIDÉO DÉTENUS PAR LES ÉTABLISSEMENTS PUBLICS ET PRIVÉS DU SECOND DEGRÉ AU 1er JANVIER 1987 (France métropolitaine)

	Établissements publics					Établ.
	Collèges	Lycées	LP	EREA	Total	privés
Audiovisuel :						
Projecteur de diapositives	41 012	12 474	6 355	259	60 100	6 818
Épiscope ou épidiascope	4 906	2 362	1 637	52	8 957	1 144
Rétroprojecteur	14 512	9 439	12 566	149	36 666	5 316
Projecteurs de cinéma tous types	13 166	3 889	1 503	92	18 650	1 554
Caméra super 8 mm	627	213	85	13	938	140
Récepteur radio	3 614	786	410	63	4 873	362
Appareil radio-cassettes	1 899	437	229	60	2 625	1 504
Magnétophones tous types	55 416	15 991	5 966	353	77 726	10 696
Électrophone et platine disque vinyl	22 659	5 453	2 260	252	30 624	4 771
Platine disque compact	216	111	25	3	355	54
Lecteurs tous types	187	119	161	1	468	49
Cabines de labo langues	2 097	3 914	350	15	6 376	1 804
Appareils photo tous types	2 300	395	367	132	3 194	799
Laboratoires de photo tous types	957	178	159	43	1 337	386
Table de mixage son	401	70	51	12	534	117
Total	**163 969**	**55 831**	**32 124**	**1 499**	**253 423**	**35 514**

Extrait de : REPERES et REFERENCES STATISTIQUES sur les enseignements et la formation, Ministère de l'Education Nationale, 1989, p.67

Le tableau des effectifs est un tableau de 15 lignes et de 5 colonnes. On dira pour simplifier que l'on analyse une situation (15x5).

b - Formalisation de la situation p x q

Les N individus d'une population sont interrogés ou observés selon deux caractères nominaux relevés simultanément.

Notons p le nombre dc modalités du caractère A, q celui de B et :

$a_1, a_2, ... a_p$ l'ensemble des modalités du caractère A

$b_1, b_2, ... b_q$ l'ensemble des modalités du caractère B.

Les individus se répartissent globalement en pq classes ou groupes. Les deux tris-à-plat donnant la répartition des individus selon chaque caractère ne permettent pas de savoir comment se répartissent les individus selon ces pq groupes.

c - Tri-croisé et tableau de contingence

Un tri-croisé (d'ordre 2) est l'opération qui consiste à lire les N fiches ou enregistrements relatifs aux N individus et à dénombrer les individus qui appartiennent à chacune des pq classes, c'est-à-dire à élaborer un tableau du type suivant :

		B					
		b_1	b_2	...	b_j	... b_q	total
A	a_1	n_{11}	n_{12}	...	n_{1j}	.. n_{1q}	$n_{1.}$
	a_2	n_{21}	n_{22}	...	n_{2j}	.. n_{2q}	$n_{2.}$
	..	..	..		..	..	
	a_i	n_{i1}	n_{i2}	...	n_{ij}	.. n_{iq}	$n_{i.}$
	..	..	..		..	..	
	a_p	n_{p1}	n_{p2}	...	n_{pj}	.. n_{pq}	$n_{p.}$
	total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	.. $n_{.q}$	

Ce tableau rectangulaire "à double-entrée" porte le nom de *tableau d'effectifs* ou *table de contingence* d'ordre 2, formé de pq "cellules". On dira que l'on est en présence d'une situation p×q.

La dernière colonne et la dernière ligne constituent les *marges* du tableau. Elles contiennent les *effectifs marginaux* - sommes correspondant aux lignes et aux colonnes de ce tableau. Ces nombres sont ceux obtenus par tri-à-plat selon chaque caractère.

On utilise ici la notation usuelle qui consiste à remplacer un des indices (de ligne ou de colonne) par un point, lorsque l'on fait la somme de tous les termes relatifs à cet indice:

$$n_{.j} = \sum_i n_{ij} \qquad \text{et} \qquad n_{i.} = \sum_j n_{ij}$$

d - Transformations du tableau de contingence

Le tableau des effectifs peut être transformé pour permettre de mieux dégager des informations, certes contenues dans le tableau initial, mais souvent masquées par l'abondance des nombres.

Plusieurs transformations sont possibles selon le type d'information que l'on recherche.

a) tableau de contingence en pourcentage

Une transformation simple consiste à remplacer le tableau des effectifs par celui des proportions, en divisant tous les nombres par l'effectif total :

proportion observée du couple (a_i,b_j) : $f_{ij} = n_{ij}/N$

proportion "marginale" de la modalité a_i :

$$f_{i.} = n_{i.}/N$$

proportion "marginale" de la modalité b_j :

$$f_{.j} = n_{.j}/N$$

Si toutes ces proportions sont multipliées par 100, on obtient un tableau en pourcentages. Des remarques analogues à celles que l'on a faites pour un seul caractère peuvent être rappelées ici :

- on gagne en lisibilité car c'est une habitude de "jauger" par rapport à un effectif total de 100 ;

- on perd toutefois de l'information et le tableau laisse à penser que la population était formée de 100 individus. Des dangers de telles transformations apparaîtront par exemple lors de la comparaison de populations (voir ci-dessous les effets dits "de structure").

b) Profils en ligne et profils en colonne

Une autre transformation intéressante consiste à faire le calcul des proportions (ou des pourcentages) "en ligne" ou "en colonne".

On appelle *profil de la modalité a_i selon le caractère B* ou plus simplement *profil de la ligne i*, la répartition en proportion ou pourcentage, selon le caractère B, des n_i individus qui possèdent la modalité a_i, c'est-à-dire la répartition:

Modalités	b_1	b_2	..	b_k
Profil de a_i	$\frac{n_{i1}}{n_{i.}}$	$\frac{n_{i2}}{n_{i.}}$	..	$\frac{n_{ik}}{n_{i.}}$

A la modalité a_i on associe la suite des "proportions conditionnelles". En pratique ces proportions sont multipliées par 100 afin de parler en pourcentages.

De façon analogue, on peut déterminer p profils "ligne", relatifs aux p modalités de A, et q profils "colonne" relatifs aux q modalités de B.

Ces profils permettent en général des comparaisons entre les divers groupes d'individus définis par les modalités d'un caractère structurel.

De plus, ils jouent des rôles essentiels dans des traitements plus sophistiqués des tableaux de contingence (par exemple l'analyse factorielle des correspondances).

CHAPITRE 3

LIEN ENTRE DEUX CARACTERES NOMINAUX

(analyse bivariée nominale)

3.1. Situations d'indépendance et de dépendance totale entre deux caractères

Indépendance

Sélectionnons deux questions nominales d'une enquête soumise à N personnes et considérons le tableau de contingence T obtenu par tri-croisé des deux caractères.
L'effectif correspondant à la cellule (i,j) est noté n_{ij}.

Pour fixer les idées, supposons que l'enquête a été effectuée auprés d'un groupe de N = 1000 téléspectateurs auxquels on a posé les deux questions suivantes (les modalités ont été arbitrairement simplifiées):

A: "*préférez-vous regarder le journal télévisé sur:* TF1 - A2 - FR3"

B: "*votre situation familiale est* : marié - Séparé ou veuf - célibataire-"

Le choix de la chaîne est indépendant de la situation familiale si on trouve la même proportion de téléspectateurs qui préférent TF1 (ou A2 ou FR3) parmi les "mariés", les "séparés" ou les "célibataires".

Cette proportion commune est alors celle du groupe étudié, toute situation familiale confondue.

De même pour ceux qui préférent A2, ou FR3.

C'est le cas si les 1000 téléspectateurs se répartissent selon un tableau de la forme:

	TF1	A2	FR3	
marié	261	232	87	580
séparé	54	48	18	120
célib.	135	12	45	300
Totaux	450	400	150	1000

Dans chaque catégorie, les pourcentages sont de:

45 % pour TF1
40 % pour A2
15 % pour FR3

Formellement, avec les notations du chapitre 2, on a pour toutes les cellules (i,j) du tableau:

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N} \qquad \text{ou encore} \qquad n_{ij} = \frac{n_{i.}\, n_{.j}}{N}$$

Un tableau de ce type, qui vérifierait strictement toutes ces relations, serait exceptionnel. Il fournirait une situation d'indépendance "idéale" et les deux caractères seraient dits *strictement indépendants*.

En pratique les effectifs observés n_{ij} ne vérifient jamais ces relations, mais s'ils s'éloignent peu de ces effectifs "théoriques" $(n_{i.}\, n_{.j})/N$ -jusqu'à une limite qu'il conviendra de préciser- on les dira *indépendants*.

Dépendance totale

A l'opposé de l'indépendance, deux caractères peuvent être *strictement associés* si, connaissant la réponse d'un individu à l'une des questions, on connaît avec certitude sa réponse à l'autre.

Ce serait le cas si le tri-croisé des deux caractères conduisait à un tableau de la forme:

	b_1	b_2	b_3
a_1	500	0	0
a_2	0	0	120
a_3	0	300	0

(si une personne a répondu la modalité a_3 à la première question, elle a nécessairement répondu b_2 à la seconde question).

Ceci exige que les deux questions aient le même nombre de modalités et que le tableau croisé des effectifs ne possède qu'une cellule par ligne et par colonne qui soit d'effectif non nul.

On dira que les deux caractères sont en association ou en *dépendance stricte* ou que le "lien" des deux caractères est maximum.

Une situation proche de celle-ci est celle où la connaissance de la réponse d'un individu à l'une des questions entraînerait la connaissance de sa réponse à l'autre, la réciproque étant fausse.

C'est le cas pour le tableau (3 x 4) :

	b_1	b_2	b_3	b_4
a_1	400	0	0	0
a_2	0	120	0	85
a_3	0	0	195	0

Si on connaît la réponse d'un individu à la question B, on en déduit sa réponse à A (sans avoir à lui poser la question); mais si on connaît sa réponse à A, un doute subsiste sur sa réponse à la question B dans le cas où sa réponse est a_2.

L'association, dans ce cas, entre les deux caractères, est totale.

En d'autres termes, dans le premier cas, les deux caractères nominaux induisent la même partition sur l'ensemble des individus; dans le second cas, la partition induite par B est "plus fine" que celle induite par A.

3.2. Mesure de l'association : le khi-deux

Entre les deux extrêmes cités ci-dessus, il y a tous les cas intermédiaires que l'on rencontre dans la réalité: les deux caractères sont plus ou moins liés.

Différentes tentatives ont eu lieu pour mesurer quantitativement ce degré de liaison, ainsi que la capacité de prévoir la réponse à une question lorsque l'on connaît la réponse à l'autre.

Les indices les plus couramment utilisés sont ceux qui mesurent la "dépendance": plus cet indice est élevé, plus le lien est étroit entre les deux caractères et meilleure est la prévision d'un caractère par l'autre

L'indice de base est celui traditionnellement appelé le "khi2" (χ^2) construit en comparant le tableau de contingence observé T et le tableau de contingence T ' ayant les mêmes marges que T mais vérifiant de plus les conditions d'indépendance stricte.

Rappelons que si le tableau T ' correspond à une telle situation d'indépendance stricte, l'effectif n'_{ij} de la cellule (i,j) vérifie :

$$n'_{ij} = \frac{n_{i.}\ n_{.j}}{N}$$

Pour chaque cellule (i,j) l'écart entre les deux tableaux est égal à

$$e_{ij} = n_{ij} - n'_{ij}$$

(remarquons que la somme de tous ces écarts est nulle).

L'indice χ^2 est construit à partir de tous ces écarts selon la formule:

$$\chi^2 = \sum_{ij} \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

(la sommation portant sur toutes les cellules du tableau de contingence).

Propriétés

1 - Cet indice est toujours positif ou nul. S'il est nul, tous les écarts sont nuls et donc les deux variables sont strictement indépendantes.

2 - Si on multiplie tous les effectifs par une constante k, le χ^2 est multiplié aussi par k. C'est donc une fonction linéaire de l'effectif total de la population interrogée.

3 - Pour un tableau de dimension (pxq), on démontre (voir l'annexe B) que le χ^2 maximum (mesure dans le cas d'une liaison maximum) est inférieur à:

$$n \,.\, \min(p-1,q-1) = n \,.\{\min(p,q) - 1\}$$

et donc aussi a fortiori à:

$$n \,.\, \sqrt{(p-1)(q-1)}$$

En notant $\nu = (p-1)(q-1)$ (nombre de degrés de liberté du tableau de contingence), et $m_{pq} = \min(p,q)$, le χ^2 vérifie donc:

$$0 \leqslant \chi^2 \leqslant n\sqrt{\nu} \quad \text{et}$$

$$0 \leqslant \chi^2 \leqslant n \,.\, (m_{pq} - 1)$$

4- En développant l'expression du χ^2, on peut le calculer par la formule:

$$\chi^2 = n.\left(\sum_{ij} \frac{n_{ij}}{n_{i.}n_{.j}} - 1\right)^2$$

qui permet parfois un calcul plus facile du χ^2.

Seuil de dépendance

En statistique inférentielle, le χ^2 est souvent utilisé comme indicateur d'indépendance de deux caractères: ayant choisi un risque d'erreur α, on détermine un seuil s tel qu'on admettra l'indépendance si le χ^2 du tableau est inférieur à s.

Dans le cas d'une étude exhaustive ou quasi-exhaustive d'une population, comme aucun modèle de nature probabiliste n'est défini pour expliquer les données observées, il ne s'agit pas du "test d'indépendance du χ^2 " .

Cependant par analogie avec la situation probabiliste, le problème consiste à trouver, pour chaque type de tableau de contingence, un seuil s tel que:

- si le χ^2 calculé est inférieur à ce seuil s, on dira que les deux caractères sont indépendants
- si le χ^2 est supérieur à s, on dira que les deux caractères sont dépendants.

Du fait d'absence de modèle, le choix de ce seuil est arbitraire. Toutefois, pour chaque degré de liberté, on choisira le seuil en s'inspirant de celui déterminé dans un test du χ^2 au risque de $\alpha = 5\%$. Ces valeurs peuvent être lues dans une "table du χ^2". Pour des degrés de liberté (ddl) inférieurs à 10, ces seuils sont:

ddl	seuil
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92

D'une manière générale, on peut prendre comme valeur approchée du seuil:

$3\sqrt{\nu} + \nu$ (où ν est le degré de liberté)

Limites d'utilisation

Le but du calcul d'un indicateur de dépendance est de comparer des couples de variables. Dans une enquête, il permettra par exemple de trouver les deux variables les plus ressemblantes...

Dans cette optique, l'utilisation du χ^2 est limitée car:

- le χ^2 dépend de l'effectif total N ; d'où l'impossibilité de comparer des couples de variables étudiées sur deux populations ou échantillons de tailles différentes.
- le χ^2 dépend de p et q ; d'où l'impossibilité de comparer des couples de variables d'une même enquête dès lors que tous les caractères n'ont pas le même nombre de modalités.

De ce fait le χ^2, qui joue par ailleurs un rôle très important dans le "test d'indépendance" en statistique inférentielle, n'est pas très performant comme indicateur de liaison. On lui préférera d'autres indicateurs déduits de lui et qui pallient les inconvénients signalés.

3.3. Mesures d'association déduites du khi-deux

Le φ^2 de Pearson ou "lien"

Pour obtenir un indicateur indépendant de la taille de la population, on peut adopter l'indicateur appelé "le φ^2 de Pearson", qui mesure aussi la dépendance entre les deux caractères A et B:

$$\varphi^2 = \frac{\chi^2}{N} \qquad \text{avec} \qquad 0 \leqslant \varphi^2 \leqslant m_{pq} - 1$$

Cet indicateur est appelé aussi le lien entre les deux caractères nominaux :

$$\varphi^2 = \text{Lien}(A,B)$$

C'est l'indicateur du χ^2 normalisé, c'est-à-dire calculé sur le tableau des proportions déduit du tableau de contingence.

Les coefficients de contingence

Pour obtenir un coefficient toujours positif et inférieur à 1, on peut utiliser l'indice noté C et appelé "coefficient de contingence de Pearson" déduit du χ^2 :

$$C = \sqrt{\frac{\chi^2}{\chi^2+N}} = \sqrt{\frac{\varphi^2}{\varphi^2+1}}$$

La fonction $x \dashrightarrow \sqrt{\frac{x}{x+1}}$ étant croissante sur R_+, le maximum de C est obtenu pour le maximum de χ^2, soit

$$\max(C) = \sqrt{\frac{m_{pq} - 1}{m_{pq}}}$$

Ainsi pour un tableau	2x2,	le maximum est	0,707
	3x4,	...	0,81
	4x4,	...	0,87

Le coefficient de contingence normé

Pour obtenir un coefficient qui soit toujours compris entre 0 (variables strictement indépendantes) et 1 (variables totalement liées), on peut choisir le coefficient de contingence normé défini par:

$$C_N = \frac{C}{\max(C)} \qquad \text{avec} \quad 0 \leqslant C_N \leqslant 1$$

Coefficient de Tschuprow

Un autre coefficient qui vérifie cette propriété d'être toujours compris entre 0 et 1, est le coefficient noté T et proposé par Tschuprow défini par :

$$T^2 = \frac{\varphi^2}{\sqrt{\nu}} \qquad \text{et on a} \quad 0 \leqslant T \leqslant 1$$

(où ν est le degré de liberté égal au produit (p-1)(q-1))

Le maximum T= 1 est atteint uniquement pour les couples de caractères ayant le même nombre de modalités (p=q) et en association stricte (cf 3.1).

Coefficient C de Cramer

Citons enfin le coefficient dit de Cramer qui vérifie encore cette propriété d'être compris entre 0 et 1, et qui est défini par:

$$C^2 = \frac{\varphi^2}{m_{pq} - 1}$$

Notons que le coefficient C de Cramer a l'avantage sur celui de Tschuprow d'être égal à 1 même si l'association n'est qu'optimum (p et q différents).

3.4. Mesures liées à la théorie de l'information

En théorie de l'information (cf Annexe C), on définit une grandeur appelée "information mutuelle de deux caractères nominaux A et B" par l'expression:

$$I(A,B) = \sum_{ij} f_{ij} \log \frac{f_{ij}}{f_{i.} f_{.j}}$$

(où f_{ij} , $f_{i.}$, $f_{.j}$ sont les fréquences jointes et marginales du tableau de contingence - la somme portant sur toutes les cellules de ce tableau).

- Cette quantité mesure le "gain" d'information réalisé lorsque l'on passe de la connaissance des marges ($f_{i.}$) et ($f_{.j}$) à la connaisance des valeurs (f_{ij}) du tableau.

- Lorsque les deux caractères sont strictement indépendants, pour toute cellule (i,j):

$$f_{ij} = f_{i.} f_{.j}$$

ce qui entraîne que tous les termes de la somme sont nuls, et donc que $I(A,B) = 0$.

- Si les deux caractères sont "voisins" de l'indépendance, un calcul d'approximation conduit à:

$$I(A,B) = \frac{1}{2 \mathrm{Ln} 2} \mathrm{Lien}(A,B)$$

et donc la mesure de l'information mutuelle équivaut, à un coefficient près, à celle du lien entre A et B, c'est-à-dire au χ^2.

- Dans le cas de la dépendance des deux caractères, les indicateurs de l'information mutuelle et du lien donnent des mesures de dépendance différentes. Toutefois, après normalisation, on obtient des indicateurs équivalents.

3.5. Cas de deux caractères dichotomiques

Il est intéressant de considérer à part le cas particulier obtenu lorsque les deux caractères nominaux sont dichotomiques, c'est-à-dire ne possèdent chacun que deux modalités (oui/non, présence/absence...).

Rappelons qu'un tel caractère A ayant pour modalités a_1 et a_2 définit une partition de la population P en deux sous-populations A_1 et A_2 telles que tout individu de A_1 admette la modalité a_1 et tout individu de A_2 admette la modalité a_2. Le tri-croisé des deux caractères A et B conduit, comme on l'a vu ci-dessus, à un tableau de contingence de la forme:

	b_1	b_2	
a_1	s	u	s+u=n
a_2	v	t	v+t=N-n
Totaux	s+v=m	u+t=N-n	N=u+v+s+t

Par codage, les modalités de tout caractère dichotomique peuvent être codées 1 (ou modalité "présence") et 0 (ou modalité "absence"). Dans un tableau de ce type le nombre s sera, de ce fait, appelé "nombre de coprésences" (ou nombre de (1,1)), et (u+v) le "nombre de discordances".

Mesure du Khi-2

En fonction des notations ci-dessus, un calcul simple permet d'écrire l'indicateur χ^2 sous la forme:

$$\chi^2 = \frac{N.\ (st-uv)^2}{(s+u)\,(v+t)\,(s+v)\,(u+t)}$$

NB : Il est intéressant de noter que si l'on considère les codes 0 et 1 comme des quantités numériques, accordant ainsi aux caractères A et B le statut de caractères numériques, le coefficient de corrélation linéaire r(A,B)=r est tel que :

$$\chi^2 = N.\ r^2$$

Hypothèse d'absence de lien

Un indicateur (proposé par I.C. Lerman) permettant de "tester" l'indépendance de deux caractères dichotomiques peut servir aussi à mesurer le lien entre les deux caractères.

Pour définir cet indicateur, on raisonne de la façon suivante:

On suppose connues:

- les réponses des N individus à la première question. On connaît donc le nombre n ainsi que les numéros d'ordre des individus parmi les N qui possèdent la modalité "1" pour le caractère A.

- le nombre d'individus m qui possèdent la modalité "1" pour le caractère B (et donc aussi le nombre N-m des individus qui ont la modalité "0" pour ce caractère).

On imagine que l'on tire au hasard (tirage sans remise) les m individus parmi les N individus de la population. auxquels on attribue la modalité "1" pour B. On s'intéresse alors au nombre aléatoire S de coprésences.

Ce problème est identique à celui qui consiste à tirer, sans remise, d'une urne formée de n boules blanches et de N-n boules noires un échantillon de m boules, et à chercher la probabilité que, parmi ces m boules, il y ait s boules blanches.

Ce nombre de coprésences est celui qui pourrait être obtenu dans le cas de "l'hypothèse d'absence de lien". Les deux caractères sont d'autant plus "liés" que le nombre de coprésences observées sera improbable par rapport à cette hypothèse d'absence de lien entre eux.

Plus précisément, la variable aléatoire (v.a.) S égale au nombre de coprésences, dans le cas d'absence de lien, suit une loi de probabilité connue sous le nom de "loi hypergéométrique" H(N,n,m).

Dans le cas où le nombre d'individus (ou de boules dans l'urne) est suffisamment grand, on admet que S suit approximativement une loi de Laplace-Gauss (ou loi normale)

de moyenne : $m(S) = nm/N$

de variance:

$$Var(S) = \frac{n(N-n)m(N-m)}{N^2(N-1)}$$

L'indice de "proximité" que propose Lerman entre les caractères dichotomiques A et B est la probabilité que la v.a S prenne une valeur inférieure à s, nombre de coprésences réellement observées. En d'autres termes les deux caractères A et B sont jugés d'autant plus voisins (ou liés) que la valeur du nombre d'individus possédant simultanément la modalité "1" est invraisemblablement grande par rapport à l'hypothèse d'absence de lien.

CHAPITRE 4

ETUDE SIMULTANEE DE TROIS CARACTERES NOMINAUX

4.1. Exemple

Dès que l'on sélectionne trois caractères nominaux pour une étude conjointe, la première difficulté provient de l'impossibilité de "voir" le tableau des effectifs croisés sous la forme naturelle qui serait un tableau tridimensionnel (à trois entrées). On est conduit à représenter les observations sous la forme d'un tableau à deux entrées, et à utiliser une subdivision de l'une des entrées pour la troisième variable.

Pour illustrer la situation, considérons les statistiques rectorales concernant les "élèves de nationalité étrangère dans le second degré de l'Académie de Rouen en 1983/84"

Sur la population des n=163.778 élèves du second degré de l'Académie, les trois caractères nominaux retenus sont:

- le secteur (Public/Privé)
- le niveau selon quatre modalités:
  - premier cycle (PC),
  - Cppn-Cpa (CPPN),
  - second cycle court (SCC),
  - second cycle long (SCL).
- la nationalité (Etrangers, Français)

Une telle situation où intervient:

un caractère nominal à 2 modalités (secteur)
un caractère nominal à 4 modalités (niveau)
un caractère nominal à 2 modalités (nationalité)

sera dite de type (2x4x2) ou "situation 2x4x2 "

Selon ces trois critères, la répartition en effectifs est donnée dans le tableau suivant:

		PC	CPPN	SCC	SCL
PUBLIC	Etrangers	2875	365	884	440
	Français	83526	7788	22091	24992
PRIVE	Etrangers	127	10	170	136
	Français	13680	596	5400	5705

Un certain nombre de tableaux d'effectifs à deux dimensions (tableaux croisés ou de contingence d'ordre deux) peuvent être créés à partir de ce tableau d'effectifs, en croisant successivement:

- secteur x niveau
  - pour les étrangers seuls
  - pour les Français seuls
  - pour Français + étrangers
- secteur x nationalité
  - pour chacun des niveaux
  - ou pour tout regroupement de niveaux (exemple: CPPN+SCC)

  (dans le cas de 4 niveaux, le nombre de tableaux de ce type serait de 15 !)
- niveau x nationalité
  - pour le public seul
  - pour le privé seul
  - pour public+privé

Dans cet exemple de situation 2x4x2, on pourrait ainsi éditer 21 tableaux de contingence d'ordre 2.

Chacun de ces tableaux peut apporter une certaine information, selon l'objectif fixé par l'étude. Mais ils ne sont pour autant pas tous aussi intéressants. Il sera donc utile de sélectionner les tableaux croisés d'ordre 2 les plus pertinents pour l'étude.

Tableaux en pourcentages ou en fréquences

Chacun de ces tableaux croisés d'ordre 2 peut être transformé en "tableau de fréquences". Selon que l'on calcule les fréquences par rapport à l'effectif global ou par rapport à l'effectif correspondant au tableau d'ordre 2 considéré, on créera des tableaux de fréquences marginales ou des tableaux de fréquences conditionnelles.

Le tableau initial peut lui aussi être transformé en tableau de fréquences (ou en pourcentages).

On désignera par (i,j,k) la cellule ou la case du tableau à trois dimensions où le symbole

i représente une modalité du premier caractère
j une modalité du deuxième caractère
k une modalité du troisième caractère

L'effectif de la cellule (i,j,k) est notée n_{ijk} et l'effectif total n. La fréquence absolue de la cellule (i,j,k) est :

$$f_{ijk} = \frac{n_{ijk}}{n}$$

On utilisera par la suite des notations semblables à celles introduites dans le cas de deux dimensions, à savoir:

$f_{ij.}$, $f_{i.k}$, $f_{.jk}$ pour les sommes des fréquences f_{ijk} portant sur toutes les modalités du caractère correspondant à la position du point

$f_{i..}$, $f_{.j.}$, $f_{..k}$ pour les sommes des fréquences f_{ijk} relativement aux deux caractères correspondant aux deux points.

(avec cette notation $f_{...}$ serait égal à 1)

Le tableau des $f_{ij.}$ correspond au tableau (à deux dimensions) des fréquences marginales entre les deux premiers caractères, toutes modalités du troisième caractère étant confondues.

Les différentes valeurs $f_{i..}$ (tableau à une seule dimension) donnent la répartition en fréquences du premier caractère. De même les $f_{.j.}$ et $f_{..k}$ donnent les répartitions des deux autres caractères.

4.2. Problèmes d'indépendance

Dès que l'on aborde les problèmes de l'indépendance ou de la liaison entre trois caractères, il est nécessaire de préciser les définitions et le type de relation recherchée.

Soient A,B,C trois caractères nominaux avec, respectivement, p,q et r modalités notées (a_i), (b_j) et (c_k).

On notera p_i la proportion des individus de type (a_i)

q_j la proportion des individus de type (b_j)

r_k la proportion des individus de type (c_k)

(p_i est donc égal à la valeur $f_{i..}$ définie ci-dessus)

4.2.1. Indépendance globale de trois caractères

Les trois caractères A, B, C sont dits *globalement indépendants* (ou indépendants dans leur ensemble) si pour toute cellule (i,j,k) la fréquence f_{ijk} (c'est-à-dire la proportion des individus ayant simultanément les modalités a_i, b_j et c_k) est égale au produit:

$$f_{ijk} = p_i.q_j.r_k$$

En pratique, sur une population réelle, le tableau des fréquences ne vérifiera jamais cette relation pour toutes les cellules. On admettra que les trois caractères sont quasi-indépendants dans leur ensemble si pour chaque cellule la fréquence observée ne s'éloigne pas trop de la fréquence théorique obtenue en faisant le produit des trois fréquences marginales.

Le tableau relatif à une situation (3x3x2) ci-dessous donne un exemple d'indépendance globale:

	b_1		b_2		b_3	
	c_1	c_2	c_1	c_2	c_1	c_2
a_1	40	160	24	96	16	64
a_2	30	120	18	72	12	48
a_3	30	120	18	72	12	48

On peut aisément vérifier que si (A,B,C) sont globalement indépendants, alors:

A et B sont indépendants
A et C sont indépendants
B et C sont indépendants

On pourra par exemple construire les tableaux croisés d'ordre 2 et vérifier qu'il en est bien ainsi.

Compte-tenu de cette remarque, une définition moins stricte de l'indépendance peut être formulée:

4.2.2. Indépendance "deux-à-deux" de trois caractères

Les trois caractères A, B et C sont dits "indépendants deux-à-deux" s'ils vérifient simultanément les trois propriétés d'indépendance de A et B, de A et C, et de B et C.

Cette condition est strictement moins forte que la condition d'indépendance globale, comme le montre l'exemple ci-dessous dans une situation (3x3x2):

	b_1		b_2		b_3		
	c_1	c_2	c_1	c_2	c_1	c_2	
a_1	20	180	34	86	26	54	400
a_2	50	100	8	82	2	58	300
a_3	30	120	18	72	12	48	300
totaux	100	400	60	240	40	160	1000

ou plus clairement encore dans la situation (2x2x2):

	b_1		b_2		
	c_1	c_2	c_1	c_2	
a_1	0	25	25	0	50
a_1	25	0	0	25	50
totaux	25	25	25	25	100

L'élaboration des trois tableaux d'ordre 2 permet de vérifier l'indépendance des trois caractères pris "deux-à-deux" alors même que les trois caractères sont globalement dépendants. En d'autres termes, on peut avoir la configuration suivante:

A est indépendant de C seul
A est indépendant de B seul
A est dépendant du couple (B,C)

4.2.3 Indépendance et dépendance conditionnelle

Il est possible aussi de fixer une des modalités de l'un des trois caractères et de se limiter à l'étude de la sous-population formée des individus possédant cette modalité.

Supposons pour fixer les idées que dans le premier exemple ci-dessus on se limite aux 200 individus ayant la modalités c_1. Ceux-ci se répartissent, relativement aux caractères A et B, selon le tableau de contingence d'ordre 2 suivant:

	b_1	b_2	b_3	
a_1	20	34	26	80
a_2	50	8	2	60
a_3	30	18	12	50
Totaux	100	60	40	200

Un calcul élémentaire permet de voir que les caractères A et B étudiés sur la sous-population de type c_1 sont dépendants. La même conclusion serait faite si on se limite aux individus ayant la modalité c_2.

En résumé, l'étude conjointe des trois caractères A, B et C permet de conclure simultanément que:

A et B sont indépendants sur l'ensemble de tous les individus (sans distinction en type c_1 ou c_2),

A et B sont dépendants sur la sous-population "c_1" (ou conditionnellement à c_1),

A et B sont dépendants sur la sous-population "c_2" (ou conditionnellement à c_2).

De façon symétrique, il est possible que les données permettent de conclure simultanément que:

A et B sont dépendants sur l'ensemble de tous les individus (sans distinction en type c_1 ou c_2),
A et B sont indépendants sur la sous-population "c_1" (ou conditionnellement à c_1),
A et B sont indépendants sur la sous-population "c_2" (ou conditionnellement à c_2).

Un exemple illustrera cette possibilité.

Sous-population "c_1"

	b_1	b_2
a_1	60	90
a_2	20	30

Sous-population "c_2"

	b_1	b_2
a_1	168	72
a_2	112	48

Ces deux tableaux vérifient la condition d'indépendance; toutefois si on les regroupe en un tableau unique croisant les deux sous-populations, la condition d'indépendance n'est plus respectée.

Cette difficulté de juger une population entière à partir de jugements partiels portés sur des sous-populations va aussi apparaitre lorsque le jugement est relatif à la comparaison de pourcentages. Il peut même conduire à des paradoxes apparents, dus à des effets de strucure.

4.3. Effet de structure

La comparaison de sous-populations en calculant les pourcentages d'individus de celles-ci possédant une certaine caractéristique peut s'avérer délicate, voire dangereuse lorsque ces sous-populations ne sont pas homogènes et sont constituées de sous-groupes au comportement différent relativement à cette caractéristique.

Illustrons cette situation par l'exemple suivant:

Il s'agit de juger l'efficacité d'un plan de formation soumis à des stagiaires pour la réussite à un test d'aptitude. Pour ce faire on compare un groupe de candidats "formés" à un groupe de candidats "témoins" qui, eux, n'ont pas suivi la formation.

On décidera que la formation est efficace si le pourcentage de réussite est plus grand parmi les candidats formés que parmi les candidats du groupe de contrôle.

Toutefois la formation de base n'étant pas la même pour tous les candidats, on prend soin de les répartir préalablement en deux catégories selon qu'ils possèdent ou non un baccalauréat, pensant que le comportement au test d'aptitude peut être différent selon que l'on est ou non bachelier.

Les 360 candidats ayant suivi l'épreuve se répartissent selon le tableau suivant:

	SANS BAC		AVEC BAC	
	Aptes	Inaptes	Aptes	Inaptes
Groupe "témoin"	21	39	75	25
Groupe "formé"	56	84	48	12

Le calcul des pourcentages de réussite dans chaque groupe de candidats conduit aux valeurs suivantes:

- pour les non-bacheliers -

- pourcentage de réussite pour les "témoins" : 35%

- pourcentage de réussite pour les "formés" : 40%

- pour les bacheliers -

- pourcentage de réussite pour les "témoins" : 75%

- pourcentage de réussite pour les "formés" : 80%

On est conduit à conclure que la formation paraît efficace autant pour les non-bacheliers (40% de réussite au lieu de 35%) que pour les bacheliers (80% au lieu de 75%) et donc que la formation est globalement efficace pour les individus soumis à l'enquête.

Cette conclusion pouvait-elle être obtenue si on ne distinguait pas les candidats selon leur formation initiale?

Regroupons les candidats, qu'ils soient bacheliers ou non. Les 360 candidats ayant suivi l'épreuve se répartissent selon le tableau suivant:

	AVEC OU SANS BAC	
	Aptes	Inaptes
Groupe "témoin"	96	64
Groupe "formé"	104	96

Le calcul des pourcentages de réussite donne alors:

- pourcentage de réussite pour les "témoins" : 60%

- pourcentage de réussite pour les "formés" : 52%

La conclusion à tirer de ces pourcentages semble en contradiction avec les conclusions précédentes sur l'efficacité de la formation puisque le pourcentage de réussite est plus important dans le groupe de contrôle que dans le groupe des stagiaires.

Cette contradiction vient du fait que les deux groupes comparés ne sont pas homogènes: il n'y a pas la même composition de candidats bacheliers et de candidats non-bacheliers. Comme la formation initiale a un effet sur la réussite au test, plus un groupe comportera de candidats bacheliers, indépendamment de la formation suivie, meilleur sera le pourcentage de réussite. Ici le groupe "témoin" est formé de 100 bacheliers sur 160 (soit un pourcentage de 62,5% de bacheliers) alors que le groupe des candidats formés comporte 60 bacheliers sur 200 (soit un pourcentage de 30% de bacheliers).

Si on ne tient pas compte de façon explicite de la formation initiale, le jugement sera faussé par cette structure différente des deux groupes. Plus que l'efficacité de la formation elle-même, c'est la plus grande aptitude de succès des bacheliers qui ressort des pourcentages de réussite.

Cette contradiction apparente dans le jugement, due à une "structure" différente des groupes sur lesquels sont calculés les pourcentages de réussite, porte le nom d'*effet de structure*.

CHAPITRE 5

ENQUETES A QUESTIONS NOMINALES

5.1. Regard critique sur un sondage

Les organes de presse, et tout particulièrement certains hebdomadaires, font fréquemment appel à des instituts de sondage pour mener des enquêtes auprès de différentes catégories de personnes : jeunes, femmes, abonnés, ...
Ils en publient ensuite les résultats partiels ou complets, souvent accompagnés d'un commentaire.

Sans vouloir porter un jugement sur l'ensemble de ces sondages, il apparaît que beaucoup d'entre eux sont analysés à partir de méthodes statistiques assez rudimentaires et que le but de l'enquête est parfois obscur.

Notre but n'est pas de discuter de la pertinence des échantillons choisis (représentativité, qualité de la collecte,...) ni de celle des questions posées -ni même de la neutralité ou non de leur formulation-, mais de regarder quels traitements statistiques ont été utilisés et quel type d'information a été retenu et publié.

Prenons un exemple parmi d'autres.

Le journal Le Monde a publié le 13 mai 1987 dans un cahier titré "Images de femmes" une enquête intitulée : "*Les femmes, comment se voient-elles ...*".

Un encart donne comme indication:

> *"Sondage réalisé par IPSOS pour Le Monde du 3 au 19 mars 1987, auprès de neuf cents femmes constituant un échantillon national représentatif de la population féminine française âgée de quinze ans et plus".*

Le résultat de ce sondage est constitué de 25 tableaux. Chacun de ces tableaux contient la répartition, en pourcentages, des femmes selon les différentes réponses possibles à une question, et proposent souvent une "ventilation" suivant un ou plusieurs critères, tels que l'âge, la région d'habitation et le type d'activité.

A titre d'exemple, la première question :

Aimez-vous ou non la manière dont on parle des femmes dans les publicités?

conduit au tableau des réponses suivant:

	aime	n'aime pas	ne se pron. pas	TOTAL
ENSEMBLE	44	44	12	100
VENTILATION				
AGE				
15-24 ans	60	32	8	100
25-44 ans	48	40	12	100
45-59 ans	39	43	18	100
60 et plus	31	57	12	100
REGION				
Paris	51	32	17	100
Province	42	46	12	100
Actives	45	44	11	100
Inactives	43	44	13	100

Quelques remarques sur un tel tableau :

- Tous les nombres publiés sont des pourcentages.
- Le nombre total des femmes interrogées est connu (900), mais on ignore la répartition des femmes par classe d'âge, par région, et par type d'activité. Il est donc impossible de transformer ces pourcentages en "nombres de femmes".

On peut bien entendu estimer ces effectifs en formulant des hypothèses complémentaires, par exemple:

> *"si le nombre de "15-24 ans" représentait le quart des 900 femmes interrogées, elles seraient environ 225. Les 32% de cette tranche d'âge qui "n'aiment pas la manière dont on montre les femmes dans la publicité" proviendraient donc de 72 réponses..."*

- Les questions correspondent:
  - soit à des caractères structurels ou critères(âge, région d'habitation...)
  - soit à des caractères conjoncturels (opinions).

Le premier type de caractère permet un éclatement de la population en sous-populations plus homogènes et donc formées de femmes dont les comportements sont a priori plus semblables. Ces caractères, sans doute utilisés à établir des "quotas" dans l'échantillon, servent ici à effectuer une "ventilation".

Le second type correspond aux caractères nominaux ; la plupart sont dichotomiques (aime/n'aime pas) - (oui/non)) - (plutôt flattée/plutôt choquée) etc..., mais la modalité supplémentaire "ne se prononce pas" est ajoutée, ce qui rend en fait tous les caractères polytomiques.

Nous observons tout d'abord que chaque question fait l'objet d'une étude séparée et que les seules opérations statistiques sont des tris-à-plat (dénombrement des femmes selon les diverses réponses proposées) et des tris-croisés avec certains critères (ventilation).

Bien entendu, ces traitements sont utiles, mais compte tenu de l'ampleur de l'information recueillie et des possibilités offertes par nombre de logiciels statistiques, il parait étonnant de se limiter à cette exploitation "au premier degré", sans chercher à connaître les relations éventuelles pouvant exister entre les caractères, ni à savoir si certains caractères sont indépendants ou non.

Notre seconde observation porte sur les textes qui accompagnent de tels sondages. Citons un commentaire de l'enquête, au sujet des 15-24 ans :

> *"ces jeunes filles ou ces jeunes femmes sont en grande majorité ravies de l'image que leur donne la publicité, moins choquées par l'érotisme et encore plus attentives à la mode que leurs mères ou leurs grands-mères. Elles sont également tentées par les vêtements masculins et la lingerie fine et elles mettent aussi plus de temps à se coiffer et à se maquiller..."*

Il est évident que le journaliste pense ainsi aider le lecteur à imaginer ce qu'est, en 1987, une jeune fille ou une jeune femme française *typique*. Il admet de façon plus ou moins implicite:

- qu'une majorité des femmes a le comportement décrit ci-dessus. Le but évident d'un tel commentaire est de définir le *portrait-type* d'une jeune fille ou d'une jeune femme de France, en 1987.
- que ce comportement décrit est celui qui est le plus *semblable* ou le plus *proche* des comportements réels (la notion de ressemblance ou de proximité n'étant ici que très vague).

Pour ce faire, le commentateur a regardé les tableaux présentés et a retenu pour chaque question, la modalité la plus souvent répondue, c'est-à-dire la valeur *modale* de chaque caractère.

Il est légitime de se poser des questions relativement à une telle démarche:

- Comment le journaliste a-t-il utilisé l'information transmise par l'institut de sondage pour rédiger ce portrait ?
- Combien de femmes vont se reconnaître dans ce portrait ?
- Est-il raisonnable d'envisager une *femme-typique* ?

- Que sont les femmes qui s'écarteraient de ce *modèle* ?

Celles qui ne se reconnaissent pas dans ce portrait sont-elles dans une catégorie *marginale*, voire *a-normale* ?

La lecture des tableaux de dépouillement de l'enquête permet d'entrevoir la méthode utilisée:

- on choisit un certain nombres de questions jugées *pertinentes*,
- on sélectionne pour chacune d'elles la modalité *majoritaire* au sens où elle a été choisie plus souvent que les autres,
- on affecte à une *femme fictive* les diverses modalités ou attributs ainsi retenus.

Nous montrons ci-dessous qu'il ne suffit pas de prendre dans chaque question la réponse "majoritaire" pour obtenir le groupe le plus important. La jeune femme décrite dans le commentaire ci-dessus peut être très "minoritaire" parmi celles de son âge ou même ne pas exister !

Quant à savoir combien de femmes se reconnaîtront dans le portrait dressé, il est impossible de le prévoir, compte tenu des informations recueillies: peut-être une majorité, peut-être quelques-unes, peut-être aucune !

5.2. Dépouillement d'un questionnaire "caractère par caractère" et profil - type

Les résultats des sondages publiés dans les médias ne sont très souvent, comme dans l'exemple ci-dessus, qu'une collection de tableaux de répartition en pourcentages correspondant aux divers caractères conjoncturels de l'enquête, jugés pertinents par le journaliste.

Ces tableaux sont relatifs soit à l'ensemble des individus, soit à des sous-populations qui correspondent à diverses modalités de certains caractères structurels. C'est la notion de *ventilation*.

Le commentaire qui accompagne souvent la publication d'une telle enquête part du principe que celle-ci permet de dégager un *profil collectif*, ou un *profil-type* c'est-à-dire de la possibilité de décrire un *individu typique* tel que les individus puissent se reconnaître plus ou moins dans celui-ci.

En d'autres termes, la méthode utilisée laisse à penser que tous les individus d'une même population, ou tout au moins que tous les individus d'une sous-population correspondant à des critères fixés -par exemple d'âge, de sexe...-, ont un comportement relativement semblable et qu'il existe de ce fait une *normalité*. C'est cette normalité que l'on fait ressortir, gommant ainsi les diversités.

La recherche de l'existence de plusieurs groupes, différenciés par leurs comportements, au sein d'une même population, n'est pas soulevée. Il existe cependant en Analyse des données des méthodes permettant d'obtenir des groupes homogènes d'individus relativement à un ensemble de variables. Ces techniques sont regroupées sous l'appellation de "méthodes de classification".

Lorsque les classes d'individus sont constituées, chacune peut alors être représentée par un élément typique, lequel dépend de la méthode de classification adoptée.

Portrait-robot et homme moyen

Le problème de l'existence d'un individu-type, que l'on pourrait appeler aussi *portrait-robot*, est identique à celui de *l'homme moyen*.

L'expression *homme moyen* date de plus d'un siècle: c'est au statisticien belge Quételet que l'on doit en 1835 une définition qui conduisit à de vives polémiques au cours des années suivantes, par exemple celle soulevée par A. Cournot en 1843, et de grandes critiques, dont celles en particulier du probabiliste J. Bertrand en 1889. A la suite de ses travaux sur la généralisation de la notion de moyenne, Fréchet réalisa en 1949 une réhabilitation de la notion statistique de l'homme moyen, à condition d'améliorer sa définition et de se protéger contre certaines aberrations.

A vrai dire, le débat concernait surtout des études faites sur des caractères numériques. Pour de tels caractères, on peut en effet calculer la moyenne des observations. L'*individu moyen*, au sens de Quételet, est alors défini comme celui qui posséderait l'ensemble des caractéristiques moyennes, même si cet individu n'existe pas, voire même ne peut exister.

Par exemple, pour définir *l'ouvrier moyen* d'une entreprise, on considère l'ouvrier fictif:

- qui a pour âge, l'âge moyen des ouvriers,
- qui touche comme salaire la moyenne des salaires,
- avec une ancienneté égale à la moyenne des années d'ancienneté,
- et avec une production égale à la moyenne des productions...

On conçoit bien que l'on puisse ainsi "créer un être difforme".

La démarche est la même pour un "portrait-robot" lorsque l'on utilise les réponses modales au lieu des moyennes.

On trouvera en annexe (C) un historique et une discussion sur le problème de l'homme moyen et des valeurs typiques.

Dangers d'aberrations

Pour illustrer le danger encouru lorsque l'on choisit la modalité "majoritaire" pour chaque question, nous prendrons deux exemples relatifs au traitement simultané de deux caractères.

Exemple 1

Soit le tableau de dépouillement (de type (3x2)) suivant:

	b_1	b_2	
a_1	31	14	45
a_2	6	34	40
a_3	15	0	15
Totaux	52	48	100

Si on utilise la méthode des modalités majoritaires, on serait tenté de dire que l'individu moyen est du type (a_1,b_1). Or en regardant le tableau croisé, on s'aperçoit qu'en fait, le groupe majoritaire est de type (a_2,b_2) (groupe formé de 34 individus) !

Exemple 2

Si on utilise la même méthode des modalités majoritaires sur le tableau de type (3x3) suivant:

	b_1	b_2	b_3	
a_1	0	20	25	45
a_2	20	12	3	35
a_3	15	2	3	20
Totaux	35	34	31	100

le profil retenu sera encore celui des individus de type (a_1,b_1) alors qu'aucun individu ne possède simultanément ces modalités !

5.3. Traitements appropriés

Les choix préalables

La réalité fourmille d'exemples d'enquêtes qui comportent un grand nombre de questions nominales (10, 50, ou 100 questions) posées à des dizaines voire des milliers d'individus. Le statisticien se trouve alors face à une telle masse d'informations qu'il est dans l'impossibilité d'extraire directement celles qui sont intéressantes, qui permettent d'apercevoir les structures sous-jacentes du phénomène étudié et de construire le modèle le plus conforme à la réalité.

Il est alors amené à faire des choix et tenter de trouver une ligne directrice, puis à soumettre les données à des traitements statistiques.

Deux choix essentiels sont à faire par le statisticien et/ou l'utilisateur, qui devra répondre préalablement aux deux questions suivantes:

- L'étude porte-t-elle sur les individus (relativement aux caractéristiques retenues) ou sur les variables (définies par leur observation sur les individus présents) ?
- Recherche-t-on des relations (ou des structures) ou une réduction-synthèse des données ?

Toute information tirée de l'étude dépendra des données brutes, mais aussi des traitements utilisés.

Etude des individus

Certaines études statistiques sont entreprises dans le but de comprendre et d'analyser l'ensemble des unités (les individus), les situer les uns par rapport aux autres.

Ce cas se présente souvent lorsque la population est observée de façon exhaustive.

Parmi les problèmes qui peuvent se poser alors, citons:

- le choix d'un individu *représentatif* parmi les individus (cas par exemple des élections);
- la définition d'un *individu-type* (cf le paragraphe précédent);
- le classement des individus selon un ordre qui tient compte du ou des caractères retenus dans l'analyse;
- la recherche de classes d'individus telles que dans chacune d'elles les individus présentent des similitudes de comportement ou d'opinion. On rassemble les individus suivant le degré de leur ressemblance ; on forme ainsi plusieurs petites familles que l'on peut encore réunir...

  Les techniques de traitement sont celles de la classification (taxinomie).

Ce dernier problème semble être le plus important dans l'optique d'un travail sur les individus. Il nécessite des choix de nature théorique, à savoir:

- définir une notion de "distance" (ou de similarité) entre les individus
- définir une méthode d'association de ces individus (indice d'agrégation pour mesurer la ressemblance entre deux classes, affectation d'un individu à un groupe...)
- définir les "qualités" d'un individu-type, représentant sa classe.

Ces choix se feront d'après l'avis de l'utilisateur qui doit avoir une idée précise sur le sens à donner à la notion de ressemblance entre individus.

Etude des variables

D'autres traitements statistiques ont pour objet les variables et les relations pouvant exister entre elles, les individus ne servant que comme support à l'étude. C'est souvent le cas d'un travail sur échantillon.

Parmi les principaux problèmes citons:

- la recherche des liaisons ou corrélations pouvant exister entre les variables (indépendances, interactions etc..), celles-ci étant souvent croisées "deux-à-deux".
- la recherche des caractères les plus pertinents (réduction du nombre des caractères) ou d'un classement des caractères par ordre d'importance.
- la recherche du ou des caractères expliquant le mieux une partition des individus (caractères discriminants).

Signalons que la recherche des différentes relations entre les variables est rendue difficile dès que leur nombre dépasse quelques unités et ceci indépendamment du nombre des individus.

Prenons, à titre d'exemple, une modeste enquête comportant 10 questions à deux modalités. Le statisticien pourra étudier chaque question isolément (soit 10 traitements et sorties); puis il peut prendre les questions "deux-à-deux" (soit 45 traitements), puis "trois-à-trois" (120 possibilités), etc...

Il devra donc faire un choix raisonné des traitements, choix qui tient compte de son expérience et des questions auxquelles il désire répondre. Il paraît illusoire de tenter tous les "croisements" possibles pour qu'apparaisse le meilleur!

Choix des tris croisés

Comme on l'a vu ci-dessus, une étape importante dans la recherche des relations entre les caractères nominaux, consiste à les croiser "deux-à-deux", et à travailler sur les tableaux de contingence.

Les caractères n'étant pas tous de même nature, il est souhaitable de distinguer trois types de tableaux selon que les caractères croisés sont structurels ou conjoncturels :

a) cas de 2 caractères structurels ou critères

Dans ce cas le tableau est simplement descriptif du choix de l'échantillon ou de la composition de la population. Il n'apporte aucune information sur le thème proprement dit de l'étude.

Dans certaines enquêtes, l'enquêteur peut fixer lui-même les effectifs marginaux de certains caractères structurels. Il choisira alors les N individus de façon à respecter ces valeurs.

On dira dans ce cas que le caractère est *à marges contrôlées* ou plus simplement *contrôlé*. Un tel caractère est appelé aussi *critère*.

On distinguera donc 3 situations différentes selon que les caractères structurels sont contrôlés ou non:

- aucun des deux caractères n'est contrôlé
- un des caractères est contrôlé, l'autre pas
- les deux caractères sont contrôlés :

  un tel tableau sert à déterminer pxq sous-populations qui respectent des contraintes d'effectifs. Mais il faut remarquer que le contrôle des marges ne suffit pas pour connaître les effectifs des pxq cellules.

Exemple :

On désire choisir un échantillon de 100 individus tel que:

- les hommes (H) et les femmes (F) soient également représentés.
- les individus soient en nombre égal dans chacune des 4 classes d'âge préalablement définies.

Donnons deux exemples, parmi beaucoup d'autres, de ventilation des individus respectant ces contraintes de marge mais conduisant à des échantillons bien différents :

Ventilation 1

	H	F	
a_1	25	0	25
a_2	25	0	25
a_3	0	25	25
a_4	0	25	25
	50	50	

Ventilation 2

	H	F	
a_1	5	20	25
a_2	12	13	25
a_3	18	7	25
a_4	15	10	25
	50	50	

b) cas d'un caractère structurel et d'un caractère conjoncturel

Il est intéressant de considérer les profils relatifs aux diverses modalités du caractère structurel et de comparer ces profils. Le caractère structurel joue le rôle de caractère explicatif pour le caractère conjoncturel.

On peut distinguer le cas où le caractère structurel est contrôlé de celui où il ne l'est pas. Dans ce dernier cas, il est possible de procéder à un *redressement* si on connait les pourcentages des diverses modalités dans la population entière.

c) cas de 2 caractères conjoncturels

Les deux caractères jouent des rôles symétriques et aucune relation de cause à effet n'est envisagée a priori.

Dans le cas d'un échantillon, avec toutefois des effectifs suffisants, on pourra "tester" l'indépendance des deux caractères.

Sinon, il est possible de mesurer le degré de liaison qui existe entre eux à l'aide d'un des indicateurs décrits plus haut.

CHAPITRE 6

PROPOSITIONS POUR UN TRAITEMENT GLOBAL

Par traitement global, nous entendons une exploitation statistique qui prend en compte simultanément l'ensemble des individus et des questions. Bien entendu, de nombreux calculs particuliers peuvent compléter le traitement global, portant sur les questions à choix multiples, les questions conditionnelles et les questions dont les modalités ne sont pas nominales.

On ne saurait trop insister sur la nécessaire distinction à respecter entre une **enquête exhaustive** et un **sondage**. L'interprétation des résultats est de nature différente et certains traitements sont spécifiques à chacune de ces situations.

Par exemple, le tri-à-plat et le tri-croisé sont utilisés pour décrire la population observée, qu'elle soit totale ou partielle.

Dans le cas d'un sondage, le tableau du tri-à-plat doit être complété, pour chacune des modalités, par un intervalle de confiance à un seuil d'erreur donné (généralement 5%), relatifs aux proportions des individus de la population entière. Pour chaque couple de variables, le tableau du tri-croisé permet de calculer le khi-deux et un test d'indépendance, basé sur la probabilité de dépassement de cette valeur, permettra de décider si l'on accepte ou non l'indépendance des deux caractères.

Dans le cas d'une étude exhaustive de la population, les indicateurs construits sur le khi-deux peuvent servir à mesurer le degré de liaison des deux caractères statistiques.

6.1. La réduction ...si nécessaire

6.1.1. Pourquoi réduire ?

Les données collectées sont parfois trop abondantes et dans cette pléthore d'informations, toutes les variables (ou toutes les questions de l'enquête) n'ont pas la même importance relativement à l'objectif particulier visé, ni la même fiabilité. L'élimination éventuelle de certaines questions et/ou de certains individus du champ d'étude a deux justifications:

1 - Réduire les temps de traitement et adapter la taille des calculs nécessaires à la capacité de l'ordinateur (le traitement global ne se conçoit qu'informatisé...).

2 - Eviter les sources de biais dues à la présence de valeurs parasites.

6.1.2. Elimination avant le traitement

a - Elimination d'individus.

La présence des individus dont les non-réponses sont trop importantes sur l'ensemble des questions faussent l'interprétation.

Le seuil d'élimination est laissée à l'appréciation de l'utilisateur. Toute élimination doit être effectuée avec prudence car :

- dans le cas d'une enquête sur toute la population statistique, l'exhaustivité n'est plus respectée. Nous avons déjà abordé ce problème en introduisant la notion de quasi-exhaustivité (cf. chap I);

- dans le cas d'un sondage, la taille de l'échantillon s'en trouve réduit. L'imprécision quant à la connaissance de la population-mère croît car les intervalles de confiance s'élargissent. C'est pour cela qu'il est conseillé de constituer un échantillon d'effectif légèrement supérieur à la taille calculée pour un degré de précision fixé.

b - Elimination de variables pour leur forme particulière.

- questions conditionnelles. Seule la question conditionnant une ou plusieurs autres (l'étage supérieur) doit être conservée. Les étages inférieurs (*Si vous avez répondu oui à la question précédente...*) doivent être exclus du fichier pour un traitement global car ils ne portent que sur une partie des individus.

- questions à choix multiples où le répondeur a pu cocher plusieurs des modalités proposées. Le total des réponses dépassent alors les 100%. Il s'agit de variables "multivoques" (au sens mathématique d'une application sur l'ensemble des parties d'un ensemble de modalités) qui seront traitées par des méthodes appropriées.

- variables purement indicatives comportant de nombreuses modalités (résultant par exemple d'une question ouverte), difficilement agrégables, où les réponses se dispersent. Le croisement systématique avec les autres variables n'aura donc qu'une signification limitée puisque ces questions n'ont pas vocation à expliquer une opinion ou un comportement.

c - Elimination des variables superflues

Ce groupe est constitué des variables hors sujet ou qui sont devenues sans objet par rapport au but de l'enquête.

6.1.3. Elimination pendant le traitement

a - Elimination des variables ayant une modalité hégémonique

Si une des modalités regroupe plus de 90% (seuil à fixer, un telle variable est alorrs constante ou quasi-constante) des individus, son aspect fortement majoritaire n'est plus de nature à induire une explication (une partition) sur les autres variables. Si sa pertinence est forte - voire triviale - dans le tri-à-plat, en revanche sa présence est génante dans le traitement global.

b - Elimination des variables aberrantes

Sous ce terme nous regroupons diverses situations courantes qui relèvent généralement d'une mauvaise formulation de la question vis-à-vis de la population concernée.

- cas où les non-réponses sont importantes : *Ne sait pas*, *Ne se prononce pas*, *Ne veut pas répondre*... (cas d'une mauvaise compréhension de la question) ou lorsque la vraie modalité est inconnue pour beaucoup d'individus;

- cas où deux variables ont des réponses jugées contradictoires. Il est préférable de traiter à part ce paradoxe - apparent ou non. Au minimum il conviendra que l'utilisateur ne retienne que l'une d'entre elles, celui-ci étant le seul juge de la pertinence ou de la cohérence des réponses à l'une ou l'autre des questions.

c - Elimination des variables quasi-dépendantes

Certaines réponses que l'on croyait différentes a priori, voire indépendantes, peuvent s'avérer -par étude du tri-croisé- très fortement dépendantes (quasi-dépendantes). Conserver dans le fichier ces deux variables est inutile puisque le lien est déjà connu. Mieux vaut n'en conserver qu'une : celle jugée la plus significative.

Ce travail de réduction opéré, le statisticien dispose de ce que nous appelons l'ensemble des variables actives.

Conclusions

Ainsi pour travailler dans de bonnes conditions, le statisticien doit-il disposer d'au moins trois fichiers:

- un fichier complet résultant du dépouillement intégral,
- un fichier ne comportant que les individus actifs pour les besoins des tris-à-plat et croisés à la demande,
- un ou plusieurs fichier(s) des individus et variables actifs permettant un traitement global, pour chacun des objectifs, dans les meilleures conditions.

6.2. Le traitement global sur les variables

Dans le cas d'un sondage, le but d'un traitement global est généralement de trouver une structure sur les variables qui repose sur la notion de dépendance ou de relation entre celles-ci. Par contre, rechercher une structure sur les individus de l'échantillon n'a pas de sens. Ce traitement sera justement réservé au cas d'une enquête exhaustive.

6.2.1. Le tri préalable entre caractères structurels et conjoncturels

On effectuera en premier lieu la distinction entre variables structurelles et conjoncturelles déjà envisagée (cf. chap I).

Ce partage est laissé à l'appréciation de l'utilisateur, du moins pour certaines d'entre elles dont le sort dépend du but de l'enquête.

6.2.2. La hiérarchie des opérations

- la première opération consiste à prendre les variables structurelles une à une et à faire le croisement systématique avec l'ensemble des variables conjoncturelles. Plusieurs indicateurs peuvent être envisagés : les sommes du khi-deux, de l'indice de Cramer, de l'indice de Tchuprow (cf. chap.3).

Cette opération permet l'obtention des variables structurelles les plus explicatives *globalement*, celles qui induisent la partition la plus "fine" ou la plus "similaire" sur les variables conjoncturelles.

Ces calculs seront d'autant moins lourds qu'auront déjà été écartées (sauf une !) les variables quasi-dépendantes entre elles tant du côté des structurelles que du côté des conjoncturelles (Cf. 6.1.)

- la deuxième opération permet d'affiner le résultat précédent. En prenant cette fois-ci chacune des variables structurelles les plus explicatives (par exemple les 3 premières dans la hiérarchie des indices) et en les croisant avec chacune des variables conjoncturelles individuellement, on trouvera - parmi les centaines de croisements possibles les caractères conjoncturels les plus dépendants de chacun des structurels retenus.

Le sens du lien sera connu : tel attribut de la population - la profession par exemple - expliquera l'opinion et non le contraire. On notera cependant que le sens de ce lien dans des cas délicats dépendra des choix opérés par l'utilisateur lors de la séparation entre variables structurelles et conjoncturelles.

- La troisième opération consiste à répéter l'opération n°1 en ne tenant compte que des variables conjoncturelles. On croise systématiquement chacune d'entre elles avec toutes les autres. On obtient alors les conjoncturelles ayant un fort lien (choix dans la liste décroissante) avec toutes les autres sans que l'on puisse déterminer le sens de ce lien puisqu'aucune n'exprime un état.

L'opinion, contrairement à l'état, est susceptible de changement ; elle n'est donc que très difficilement une variable explicative.

- La quatrième opération s'inspire de l'opération n°2. Elle opère le même travail avec un nombre de variables réduit aux seules conjoncturelles mais en les prenant deux à deux parmi celles ayant les plus forts liens sans en connaître le sens. On disposera alors des paires d'opinion les plus semblables.

6.2.3. Le classement ultime entre explicatifs et expliqués

L'opération précédente permet déjà de déceler -parmi la totalité des variables- celles qui apparaissent les plus significatives au cours du traitement global. Ce sont les variables dites quasi-dépendantes entre elles.

Cette liste de variables réduites doit faire l'objet d'un nouveau partage entre variables explicatives et expliquées.

Dans le groupe des variables explicatives figureront :

- les structurelles retenues résultant de la 2ème opération
- les conjoncturelles jugées explicatives (le sens du lien ayant été tranché par l'utilisateur -cf. 4ème opération).

Le lot des variables expliquées ne sera constitué que de conjoncturelles parmi celles sélectionnées lors de la 4ème opération mais jugées non explicatives.

Une cinquième et dernière opération cherche pour chaque variable expliquée les 2 où 3 variables explicatives qui concourent à donner le maximum de dépendance, car l'influence n'est jamais unique.

Fort de ces résultats les plus pertinents, le statisticien est alors en mesure d'en tirer une appréciation globale et de rédiger un commentaire de synthèse.

6.3. Traitement sur les individus

Dans le cas où la population est étudiée de façon exhaustive, la synthèse la plus radicale que l'on puisse faire est de donner un individu type, c'est-à-dire un nivellement total des informations récoltées sur cette population.

Pour donner une image moins grossière, on fera une classification et en particulier une hiérarchie permettra de définir des catégories d'individus aussi homogènes que possible par rapport aux caractères étudiés, une typologie...

Ces approches nécessitent une notion de similarité, de distance entre individus, puis de ressemblance entre classes ou entre individu et classe pour affecter un individu à une classe ou regrouper deux classes, suivant le dicton "qui se ressemble s'assemble".

Plusieurs choix sont possibles pour définir une distance (plus précisement un écart) entre individus, entre autres:

- une distance obtenue à partir du tableau disjonctif complet croisant individus et modalités, donnée à l'annexe D
- la distance du khi2 donnée dans l'annexe B

De même, plusieurs choix sont possibles pour mesurer la ressemblance entre deux groupes d'individus:

- la distance (écart) du lien maximum, c'est-à-dire la distance la plus grande existant entre un élément d'un groupe et un élément de l'autre;
- la moyenne des distances entre les individus pris dans chacun des groupes;
- la distance entre individus typiques de chacun des groupes;
- la distance du khi2 entre groupes.

Une fois ces choix effectués, on part de la partition la plus fine sur la population, dont les classes sont réduites à un élément. On construit une nouvelle classe en réunissant les deux classes les plus proches pour l'écart choisi, et on continue la procédure jusqu'à obtenir le nombre de classes désiré, ou jusqu'à ce que l'écart entre les classes constituées soit plus grand qu'un seuil fixé, ou jusqu'à ce que tous les individus soient réunis en une seule classe.

L'ensemble de ces partitions successives forment la *hiérarchie*.

Cette méthode est connue sous le nom de *classification hiérarchique ascendante*.

ANNEXE A

REPRESENTATIONS GRAPHIQUES

La lecture des informations contenues dans un tableau pxq est facilitée par l'élaboration de diagrammes ou de graphiques.

Les techniques de représentation sont très différentes selon les problèmes, les préoccupations ou le modèle mathématique utilisés pour formaliser les données.

Citons 3 types principaux de techniques :

- Diagrammes (par exemple en rectangles)
- Matrices permutables et matrices de type Bertin
- Méthodes euclidiennes: analyse des correspondances.

On pourra en trouver un exposé et des exemples dans la bibliographie sommaire ci-dessous.

1 - Diagrammes

Lorsque l'un des caractères est structurel et l'autre conjoncturel, le tableau de données peut être transformé en un tableau de profils (cf. le chapitre 2). Chaque profil correspond à une répartition qui peut être représentée par un graphique, appelé diagramme, constitué d'un ensemble de rectangles ayant des bases alignées et de même largeur. Si le caractère structurel a p modalités, on peut dessiner p diagrammes de ce type, associé chacun à une modalité.

2 - Matrices permutables et matrices de type Bertin

Les tableaux de contingence d'ordre 2 construits à partir de deux caractères nominaux sont des matrices "permutables" au sens qu'on peut permuter les lignes entre elles, ou les colonnes entre elles, sans dégradation de l'information: en effet, l'ordre des lignes et des colonnes résulte d'un choix arbitraire.

Le géographe J. Bertin a donné à de telles matrices le nom de *matrices ordonnables*. Un traitement graphique de ces tableaux est basé sur la création et la manipulation d'une matrice de "dominos", chaque domino ayant une valeur de gris (du noir au blanc) représentant l'effectif n_{ij} d'un groupe.

Les lignes et/ou les colonnes de ces dominos peuvent ensuite être permutées par un opérateur jusqu'à ce qu'il mette en évidence une structure (blocs disjoints, diagonale...).

3 - Analyse des correspondances

Depuis quelques années (milieu des années 60), des méthodes puissantes de représentation et d'analyse des données ont été développées et ceci grâce à l'apport de modèles mathématiques et de nouvelles possibilités de calculs offertes par la montée en puissance des (micro-)ordinateurs.

Parmi ces techniques, une a été développée particulièrement à l'Université de Paris, sous la direction du Professeur Benzécri, et est bien adaptée au traitement des tableaux de contingence. Faisant partie d'une classe plus vaste de traitements appelés "analyses factorielles", elle porte le nom d'"*analyse factorielle des correspondances*" (en abrégé AFC). Disons simplement ici qu'il s'agit de "mesurer" les distances entre les profils "ligne" (et simultanément entre profils "colonne") et de trouver des représentations planes (dans le plan) où figurent des points représentant les diverses modalités des 2 caractères A et B ou les individus.

La proximité ou l'éloignement des points permet de porter des jugements sur la population étudiée. (voir Bibliographie).

Bibliographie

Sur les diagrammes:

- CALOT G. : Cours de statistique descriptive, Dunod, 2e éd., 1973.
- GRAIS B. : Techniques statistiques, t.1: statistique descriptive, Dunod, 2e éd. 1979.

Sur l'analyse des correspondances:

- de LAGARDE : Initiation à l'analyse des données, Dunod, 1983.
- VOLLE M.: Analyse des données, Economica, 1978.

Sur les méthodes de matrices permutables:

- BROCARD : Le traitement graphique en géographie, Cahiers géographiques de Rouen, n°20, 1983.
- BERTIN : La graphique et le traitement graphique de l'information, Flammarion, 1977.
- Collectif : La graphique, guide méthodologique pour la pratique du travail autonome, CRDP Besançon, 1983.

ANNEXE B

DISTANCE DU χ^2 EN ANALYSE DES DONNEES

et REPRESENTATIONS GEOMETRIQUES

Cette annexe présente aux numéros 1 à 4 quelques éléments des notions mathématiques utilisées en analyse des données. Le lecteur non mathématicien peut laisser de côté cette approche plus théorique.

1 - FONCTIONS ET MESURES SUR UN ENSEMBLE FINI. DUALITE.

Soit $I = \{1,\ldots, k\}$ un ensemble fini. L'ensemble E des fonctions de I dans $\mathbb{R}$ peut alors être identifié à $\mathbb{R}^k$ en associant à tout élément x de E le vecteur $(x^i)_{1\le i\le k}$, où $x^i = x(i)$. La base canonique $(e_i)_{1\le i\le k}$ de cet espace vectoriel E de dimension k sur $\mathbb{R}$ est constituée des vecteurs unité des différents axes [par exemple, $e_1 = (1,0,\ldots,0)$], et tout vecteur x de E se décompose par rapport à cette base sous la forme: $x = \sum_{i=1}^{k} x^i e_i$.

Une mesure ou pondération sur I est une application $\mu : \mathcal{P}(I) \longrightarrow \mathbb{R}^+$ vérifiant $\mu(\emptyset) = 0$ et pour A et B $\subset$ I disjoints, $\mu(A\cup B) = \mu(A) + \mu(B)$.
Puisque l'ensemble I est fini, la mesure μ est complètement déterminée par le poids affecté à chaque élément i qui sera noté $\mu(\{i\}) = \mu_i$, $i\in I$.
Une pondération μ sera dite *pondération stricte* si $\forall i \in I$, $\mu_i > 0$.
Une pondération de masse totale $\mu(I) = 1$ sera dite *mesure de probabilité*, ou plus simplement probabilité.

Pour une mesure μ sur I, l'intégrale d'une fonction x sur I correspond à une somme pondérée (puisque I est fini) et s'écrit: $\mu(x) = \sum_{i\in I} \mu_i x^i$.

Dans le cas où μ est une probabilité, $\mu(x)$ est la moyenne des valeurs x^i pondérées par les poids μ_i.

L'intégrale à base μ définit une forme linéaire sur E: si α, β sont des réels et x, y des fonctions de E, on a $\mu(\alpha x + \beta y) = \alpha\mu(x) + \beta\mu(y)$.
De plus, si les valeurs de x sont positives, $\mu(x) \ge 0$.

Les formes linéaires sur E définissent un espace vectoriel E*, appelé *dual* de E , et on a la relation de dualité entre E et E*:

$\langle x^*, x\rangle = x^*(x)$, où $x^*\in E^*$ et $x\in E$.

On peut alors identifier l'ensemble des mesures sur I et l'ensemble $\mathbb{M}$ des formes linéaires positives sur E : $\mathbb{M} \subset E^*$, et dans ce cas, la relation de dualité s'écrit : $\langle \mu, x \rangle = \mu(x) = \sum_{i \in I} \mu_i x^i$.

2 - DUALITE DANS LES ESPACES EUCLIDIENS DE DIMENSION FINIE.

Le dual E^* est aussi un espace vectoriel de dimension k sur $\mathbb{R}$. A la base canonique $(e_i)_{1 \le i \le k}$ sur E, on associe la base canonique duale $(e_i^*)_{1 \le i \le k}$ sur E^* caractérisée à partir de la relation de dualité par: $\langle e_i^*, e_j \rangle = \delta_{ij}$, où $\delta_{ij} = 1$ si $i = j$ et $\delta_{ij} = 0$ si $i \neq j$ (δ_{ij} est le symbole de Kronecker).

En notation matricielle, par rapport aux bases canoniques, un élément x de E est représenté par une matrice colonne, un élément x^* de E^* par une matrice ligne et la relation de dualité entre x^* et x est le produit matriciel de la ligne par la colonne, qui est un nombre.

On désigne par φ un produit scalaire sur E, c'est-à-dire que l'application $\varphi : E \times E \longrightarrow \mathbb{R}$ est une forme bilinéaire symétrique et définie positive (non dégénérée) qui vérifie les propriétés:
si α, β sont des réels et x, y, z des éléments de E,

$\varphi(\alpha x + \beta y, z) = \alpha \varphi(x,z) + \beta \varphi(y,z)$

$\varphi(x,y) = \varphi(y,x)$ et

$\varphi(x,x) = 0 \iff x = 0$.

Un espace vectoriel E muni d'une telle structure est dit espace euclidien pour le produit scalaire φ. L'application $x \in E \mapsto \sqrt{\varphi(x,x)}$ définit une norme sur E, qu'on notera $\|x\|$.

Le produit scalaire φ sur E est entièrement déterminé par ses valeurs $\varphi_{ij} = \varphi(e_i, e_j)$, $1 \le i, j \le k$, pour les éléments de la base canonique car si:
$x = \sum_i x^i e_i$ et $y = \sum_i y^i e_i$ sont des éléments de E, alors $\varphi(x,y) = \sum_{i,j} \varphi_{ij} x^i y^j$.

Par rapport à la base canonique $(e_i)_i$ de E, on peut écrire cette relation sous forme matricielle : $\varphi(x,y) = {}^t x \, \varphi \, y$. Ici, on désigne par la même lettre le vecteur x et la matrice colonne associée ($^t x$ étant la matrice ligne de x), le produit scalaire φ et la matrice carrée $(\varphi_{ij})_{1 \le i,j \le k}$ associée à φ.

Pour $x \in E$ fixé, l'application $y \in E \mapsto \varphi(x,y)$, qu'on note $\varphi(x,.)$, est une forme linéaire sur E (un élément de E^*), qui est déterminée de façon unique par x. On définit ainsi une application $\Phi : E \longrightarrow E^*$ par $\Phi(x) = \varphi(x,.)$ qui est un isomorphisme d'espaces vectoriels de E sur E^*.

Φ devient un isomorphisme d'espaces euclidiens si l'on munit E^* du produit scalaire dual φ^* défini par: $\varphi^*(\Phi(x),\Phi(y)) = \varphi(x,y)$, ou encore :

$\varphi^*(x^*,y^*) = \varphi(\Phi^{-1}(x^*),\Phi^{-1}(y^*))$, pour x^* et y^* dans E^*.

Par cet isomorphisme, la relation de dualité entre E et E^* vérifie :

$\langle \Phi(x), y\rangle = \varphi(x,y)$ ou encore $\langle x^*, y\rangle = \varphi(\Phi^{-1}(x^*), y)$.

3 - METRIQUE DU χ^2 A BASE φ

Dans le cas où la matrice φ est diagonale: $\forall i,j,\ \varphi_{ij} = \varphi_i \delta_{ij}$, le produit scalaire prend la forme plus simple : $\varphi(x,y) = \sum_i \varphi_i x^i y^i$ et on peut considérer $(\varphi_i)_{1\le i\le k}$ comme une pondération stricte chargeant les axes canoniques. Les éléments de la base canonique sont alors orthogonaux pour le produit scalaire φ.

Réciproquement, si φ est une pondération stricte sur I, (d'où $\varphi_i \neq 0$, $\forall i \in I$), alors on peut définir sur E un produit scalaire, noté aussi φ, par: $\varphi(x,y) = \sum_i \varphi_i x^i y^i$.

Ce produit scalaire induit une norme et une distance sur E qui est une norme ou une métrique euclidienne :

$$\|x\|_\varphi^2 = \varphi(x,x) = \sum_i \varphi_i (x^i)^2 \qquad \text{et} \qquad d_\varphi(x,y) = \|x - y\|_\varphi$$

$\|x\|_\varphi^2$ apparait comme le moment d'inertie par rapport à 0 des points x^i affectés des masses φ_i.

Comme on l'a vu au n° 2, le produit scalaire φ sur E définit un isomorphisme Φ de E sur E^* et par la relation de dualité on a :

$$\langle \Phi(x), y\rangle = \varphi(x,y) \ , \ \forall\ x, y \in E.$$

En particulier, pour une forme linéaire positive, c'est-à-dire une mesure $\mu = \Phi(x) \in \mathbb{M}$, on a: $\forall i,\ \mu_i = \varphi_i x^i$; μ est la mesure de densité x par rapport à φ considéré comme pondération.

Réciproquement, si $\mu \in \mathbb{M}$, $x = \Phi^{-1}(\mu)$ est donné par $\forall i,\ x^i = \dfrac{\mu_i}{\varphi_i}$

et x est la densité de μ par rapport à φ.

On obtient le produit scalaire associé entre mesures :

$$\varphi^*(\mu,\nu) = \varphi(\Phi^{-1}(\mu),\Phi^{-1}(\nu)) = \sum_i \varphi_i \frac{\mu_i}{\varphi_i}\frac{\nu_i}{\varphi_i} = \sum_i \frac{\mu_i\ \nu_i}{\varphi_i} \quad \text{si } \mu \text{ et } \nu \in \mathbb{M},$$

et cette distance, appelée distance du χ^2 de base φ entre les deux mesures μ et ν, est donnée par :

$$d_\varphi^{\,2}(\mu,\nu) = \sum_i \frac{(\mu_i - \nu_i)^2}{\varphi_i}$$

4 - LOI MULTINOMIALE et APPROXIMATION du χ^2

On considère une variable X, définie sur une population, prenant la modalité x_i avec la fréquence p_i, pour $1 \le i \le k$.

On tire au hasard un échantillon de taille n avec remise dans cette population, et on note N_i le nombre d'individus dans l'échantillon présentant la modalité x_i. La variable multidimensionnelle $N = (N_1, \ldots, N_k)$ suit alors une loi multinomiale $\mathcal{M}(n ; p_1, \ldots, p_k)$.

En posant $X_i = \dfrac{N_i - np_i}{\sqrt{np_i}}$, la loi de la variable $\sum_i X_i^2 = \sum_i \dfrac{(N_i - np_i)^2}{np_i}$ converge, lorsque $n \to \infty$, vers une loi de probabilité appelée loi du χ^2 à k-1 degrés de liberté .

Cette quantité est aussi la distance (définie au n° précédent) entre la distribution observée $(N_i)_i$ et la distribution théorique $(np_i)_i$ en prenant pour base cette distribution théorique.

Ces considérations expliquent donc le nom "*distance du* χ^2" donné à cette distance entre mesures.

5 - χ^2 DE CONTINGENCE

On considère 2 variables nominales X et Y, définies sur une population, ayant pour ensembles de modalités $\{x_i\}_{1 \le i \le k}$ et $\{y_j\}_{1 \le j \le \ell}$.

Dans cette population (ou un échantillon) de taille n, on note n_{ij} le nombre d'individus ayant donné comme réponses x_i et y_j, pour chaque valeur de i et j.

Ces effectifs n_{ij} définissent une mesure sur l'ensemble des modalités $\{(x_i, y_j)\}_{i,j}$ du couple (X,Y) de variables. Cette mesure, qui est la répartition ou distribution du couple (X,Y), est représentée par le tableau de contingence en effectifs :

X \ Y	y_1	..	y_j	..	y_ℓ	
x_1	n_{11}	..	n_{1j}		...	$n_{1.}$
..	..					..
x_i	n_{i1}	..	n_{ij}	..		$n_{i.}$
..	..					..
x_k	n_{k1}					$n_{k.}$
	$n_{.1}$	..	$n_{.j}$		$n_{.\ell}$	n

Les répartitions $(n_{i.})_i$ et $(n_{.j})_j$ sont les marges du tableau (voir II-2.5)

En divisant tous les effectifs n_{ij} par n: $f_{ij} = \frac{n_{ij}}{n}$, on obtient la répartition en fréquences $(f_{ij})_{ij}$ et le tableau de contingence en fréquences.

Sous l'hypothèse de l'indépendance des 2 variables X et Y, les marges étant fixées, la répartition théorique est: $\nu_{ij} = \frac{n_{i.}\ n_{.j}}{n}$, $1 \le i \le k$, $1 \le j \le \ell$.

Le carré de la distance du χ^2 de base $\nu = (\nu_{ij})_{ij}$ entre répartition observée et répartition théorique est

$$C^2 = \sum_i \sum_j \frac{(n_{ij} - \nu_{ij})^2}{\nu_{ij}} \qquad \text{(dit } \chi^2 \text{ de contingence)}$$

Pour un échantillon tiré au hasard avec remise, la distribution de C^2 tend vers la loi du χ^2 à $(k - 1) \cdot (\ell - 1)$ degrés de liberté.

6 - PROFIL-LIGNE ET DISTANCE DU χ^2.

Le tableau de contingence peut être transformé pour obtenir sur chaque ligne, par exemple celle de x_i, la distribution conditionnelle de la variable Y par rapport au sous-groupe des individus ayant répondu x_i.

$X \backslash Y$	y_1	y_j	y_l	
x_1				
x_i		f_i^j		$f_{i.}$
x_k				
		$f_{.j}$		

où $f_i^j = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}$

$\forall$ i, on a $\sum_j f_i^j = 1$

On dira que (f_i^j), $1 \le j \le \ell$, est le profil de la ligne i.

C'est une répartition sur l'ensemble des modalités $\{y_j\}_j$ de la variable Y, c'est-à-dire d'après le n° 1, un élément de $(\mathbb{R}^\ell)^*$. On peut le représenter par un point P_i de l'espace $(\mathbb{R}^\ell)^*$, et lui affecter le poids $f_{i.}$ associé à la modalité x_i dans la population.

Tous les points P_i sont dans une même partie d'un hyperplan de $\mathbb{R}^{\ell *}$: le simplexe des distributions défini par : $\forall$ j, $\alpha_j \ge 0$ et $\sum_{j=1}^{\ell} \alpha_j = 1$.

La distribution marginale $(f_{.j})_{1 \le j \le \ell}$ est la moyenne pondérée par $(f_{i.})_i$ des profils-lignes : $\forall$ j = 1 à ℓ, $\sum_i f_{i.}\ f_i^j = f_{.j}$.

Le point G représentant cette répartition marginale est le centre de gravité du nuage de points pondérés $(P_i, f_{i.})_{1 \le i \le k}$.

Si l'on munit $\mathbb{R}^{\ell *}$ de la distance du χ^2 de base $(f_{.j})_j$ pour laquelle c'est un espace euclidien, le calcul de l'inertie I du nuage $(P_i, f_{i.})_i$ par rapport à son centre de gravité donne:

$$I = \sum_{i=1}^{k} f_{i.}\, d^2_{(f_{.j})}(P_i, G) = \sum_{i=1}^{k} f_{i.} \sum_{j=1}^{\ell} \frac{(f_i^j - f_{.j})^2}{f_{.j}} \quad \text{d'où} \quad I = \sum_{i,j} \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}$$

Cette valeur est aussi le coefficient φ^2 de Pearson et on a la relation avec le χ^2 de contingence calculé C^2 : $I = \varphi^2 = \frac{C^2}{n}$, où n est la taille de la population ou de l'échantillon.

7 - EFFET D'UN REGROUPEMENT DE MODALITES.

Dans le tableau de contingence, on considère 2 lignes qu'on supposera être les lignes 1 et 2, dont les points représentatifs P_1 et P_2 ont pour coordonnées dans $\mathbb{R}^{\ell}$ les profils-lignes $(f_1^j)_j$ et $(f_2^j)_j$.

Le barycentre P_0 des points P_1 et P_2 est déterminé par:

$(f_{1.} + f_{2.})\, P_0 = f_{1.} P_1 + f_{2.} P_2$ et P_0 est affecté du poids $f_0 = f_{1.} + f_{2.}$

On notera $(f_0^j)_j$ le profil-ligne correspondant à P_0, qui est donc caractérisé par : $\forall j,\ f_0^j = \frac{1}{f_{1.} + f_{2.}} (f_{1.} f_1^j + f_{2.} f_2^j)$.

En notant I l'inertie du nuage initial et I_0 l'inertie du nuage obtenu en remplaçant P_1 et P_2 par leur barycentre P_0, on obtient :

$$I - I_0 = \frac{f_{1.} f_{2.}}{f_{1.} + f_{2.}} \sum_{j=1}^{\ell} \frac{(f_1^j - f_2^j)^2}{f_{.j}} = \frac{f_{1.} f_{2.}}{f_{1.} + f_{2.}}\, d^2_{(f_{.j})}(P_1, P_2)$$

• On a d'abord $I_0 \le I$: le χ^2 de contingence diminue quand on regroupe des modalités, ce qui est équivalent de dire, dans l'interprétation mécanique, que l'inertie d'un système de masses diminue quand on remplace 2 points par leur barycentre.

• Ensuite, $I_0 = I$ si et seulement si les 2 points P_1 et P_2 sont confondus, c'est-à-dire qu'alors les profils-lignes sont identiques ; dans ce cas, le χ^2 de contingence reste constant par regroupement de ces modalités.

Ces considérations indiquent une marche à suivre pour effectuer une réduction d'un questionnaire par regroupement de modalités d'une variable Λ.
On désigne par B_k les autres variables ayant chacune m_k modalités, $1 \le k \le \ell$.

Dans les tableaux de contingence entre A et B_k, pour chaque couple $(a_i, a_{i'})$ de modalités de A, on calcule les profils des lignes i et i' ayant pour points représentatifs $P_i^k = ({}^k f_i^j)_j$ et $P_{i'}^k = ({}^k f_{i'}^j)_j$ où $1 \leq j \leq m_k$.

La distance du χ^2 entre ces points est : $d^2(P_i^k, P_{i'}^k) = \sum_{j=1}^{m_\ell} \frac{({}^k f_i^j - {}^k f_{i'}^j)^2}{f_{.j}}$

On choisira alors de regrouper les modalités a_i et $a_{i'}$ de A qui rendent la somme $\sum_k d^2(P_i^k, P_{i'}^k)$ minimum, c'est-à-dire que les modalités a_i et $a_{i'}$ sont celles qui se ressemblent le plus par rapport à toutes les autres variables.

8 - DISTANCE DU χ^2 entre INDIVIDUS

On a défini précédemment la distance du χ^2 entre deux profils-ligne obtenus à partir d'un tableau de contingence croisant deux caractères. On peut faire le même genre de travail à partir des individus.

Les données globales sont regroupées dans un tableau qui croise les individus et les variables et contient les différentes modalités des variables. On va le transformer par codage pour en obtenir un autre contenant des valeurs numériques 0 ou 1.

Un tel tableau binaire, qui croise les individus et l'ensemble $\mathbb{M}$ des modalités des variables, est appelé disjonctif complet. C'est un tableau de présence-absence sur toutes les modalités, c'est-à-dire qu'on décompose chaque question en plusieurs questions dichotomiques de la forme: "répondez-vous la modalité m" pour chaque modalité de la question.

Soit une question Q_j admettant les modalités $m_j^1, \ldots, m_j^k$, si l'individu ω_i répond la modalité m_j^r à cette question, on notera 1 dans la case correspondante du tableau et 0 pour les autres modalités de la question.

question Q_j

individus \		m_j^1 .. m_j^r .. m_j^k	
ω_i		0 .. 1 .. 0	

p

La somme des valeurs d'une ligne est égale à p, nombre de questions, et la somme des valeurs d'une colonne est égale à l'effectif des réponses de cette modalité.

On note z_i^m la valeur dans le tableau binaire pour l'individu ω_i et une modalité m. Le vecteur des réponses de l'individu ω_i sera $(z_i^m)_{m \in \mathbb{M}}$.

La distance du χ^2 entre les individus ω_i et $\omega_{i'}$ est donné par:

$$d^2_{\chi^2}(\omega_i, \omega_{i'}) = \sum_{m \in \mathbb{M}} \frac{np}{z^m_\cdot} \left[\frac{z^m_i}{z^\cdot_i} - \frac{z^m_{i'}}{z^\cdot_{i'}} \right]^2 ,$$ en supposant que toutes les modalités sont effectivement prises, c'est-à-dire que $\forall m,\ z^m_\cdot \neq 0$.

Ici, $z^m_\cdot = \sum_i z^m_i$ est l'effectif de la modalité m ; on le notera s(m), c'est le nombre de fois où la modalité m est prise dans la population.

Du fait que pour chaque question, un individu ne donne qu'une seule modalité en réponse, $\forall i,\ z^\cdot_i = \sum_m z^m_i = p$, cette formule peut se simplifier :

$$d^2_{\chi^2}(\omega_i, \omega_{i'}) = \frac{n}{p} \sum_{m \in \mathbb{M}} \frac{(z^m_i - z^m_{i'})^2}{z^m_\cdot} = \frac{n}{p} \sum_j \sum_{m \in \mathbb{M}_j} \frac{(z^m_i - z^m_{i'})^2}{z^m_\cdot} ,$$

Pour la question Q_j, si les individus ω_i et $\omega_{i'}$ donnent la même réponse, la valeur sera 0, et s'ils donnent des réponses différentes m et m', la valeur sera $\frac{1}{z^m_\cdot} + \frac{1}{z^{m'}_\cdot} = \frac{1}{s(m)} + \frac{1}{s(m')}$,

d'où $d^2_{\chi^2}(\omega_i, \omega_{i'}) = \frac{n}{p} \sum_j \left(\frac{1}{s(x^j_i)} + \frac{1}{s(x^j_{i'})} \right) \delta^j_{i,i'}$, où $s(x^j_i)$ est le nombre de fois où la modalité x^j_i, réponse de ω_i à la question Q_j, est prise dans la population, et $\delta^j_{i,i'}$ est 0 si ω_i et $\omega_{i'}$ ont même réponse, 1 pour des réponses différentes à la queestion Q_j.

La distance du χ^2 entre individus pondère en donnant moins d'importance aux modalités d'occurence fréquente.

ANNEXE C

THEORIE DE L'INFORMATION

1 . NOTION D'INFORMATION

Soit E un ensemble contenant n éléments. On se demande quel est le nombre minimum de questions dichotomiques (à réponses "oui" ou "non") pour déterminer à coup sûr un élément quelconque de E. Chacune de ces réponses "oui" ou "non" sera appelée "*unité d'information*".

a) Si $n = 2^k$, on voit facilement qu'il suffit de poser k questions de façon judicieuse, c'est-à-dire d'obtenir k unités d'information, pour déterminer un élément de E. Par exemple, en repérant chaque élément de E par un nombre codé en binaire (avec les chiffres 0 ou 1 correspondant à des réponses "oui" ou "non"), il faut et il suffit de poser k questions associées à chaque chiffre binaire du nombre.

On dit que k est la *quantité d'information* à apporter pour déterminer un élément de l'ensemble E contenant $n = 2^k$ éléments, d'où $k = \log_2 n$ (*)[1] ; c'est la définition de HARTLEY (1928). On dit aussi que k est le degré d'indétermination ou d'incertidude sur E.

b) Si l'ensemble E contient n éléments, où n n'est pas nécessairement de la forme 2^k, soit k un entier vérifiant: $2^{k-1} < n \leq 2^k$.
On peut à coup sûr déterminer un élément de E en posant k questions, mais quelquefois il en suffit de k-1. En répétant cette opération de détermination, on voit qu'en moyenne le nombre de questions à poser pour repérer un élément se rapproche de log n.

En effet, en appelant m-séquence une suite formée de m éléments de E avec répétitions, le nombre de telles suites est n^m et on repèrera à coup sûr une m-séquence avec k_m questions où : $2^{k_m - 1} < n^m \leq 2^{k_m}$, c'est-à-dire :
$k_m - 1 < \log n^m \leq k_m$ ou encore $\log n^m \leq k_m < \log n^m + 1$.

(*) Dans la suite, on écrira log pour le logarithme à base 2 et Ln pour le logarithme népérien. On a alors la relation : $\log n = [\frac{\text{Ln } n}{\text{Ln } 2}]$

Donc, le nombre moyen de questions par élément d'une m-séquence vérifiera $\log n \leq \frac{k_m}{m} < \log n + \frac{1}{m}$, d'où le résultat quand on fait croître m.

En notant $\mathcal{J}(E)$ la *quantité d'information* à apporter pour déterminer un élément de E, on a la **formule de Hartley** : $\mathcal{J}(E) = \log(\mathrm{Card}(E))$, où Card(E) est le nombre d'éléments de E.

Cette fonction $\mathcal{J}$ possède les propriétés :

1) $\mathcal{J}(E \times F) = \mathcal{J}(E) + \mathcal{J}(F)$

2) si $E \subseteq F$ alors $\mathcal{J}(E) \leq \mathcal{J}(F)$

3) $\mathcal{J}(\{0,1\}) = 1$

Ici, la propriété 3) donne l'unité de mesure de l'information ; 2) signifie que plus un ensemble est grand, plus il faut avoir d'information pour caractériser chaque élément de l'ensemble ; 1) se justifie de la façon suivante : on peut déterminer un couple (x,y) de $E \times F$ en caractérisant l'élément x de E et l'élément y de F, i.e. la quantité d'information est une grandeur additive.

Réciproquement, une fonction d'ensembles vérifiant les conditions 1,2,3) est la fonction $\mathcal{J}$.

2 . INFORMATION APPORTEE PAR UN CARACTERE.

Sur un ensemble E ayant n éléments, on considère un caractère X (c'est-à-dire une application) défini sur E et prenant ses valeurs dans l'ensemble $M_X = \{ x_i \}_{1 \leq i \leq k}$.

Ce caractère X détermine une partition $\mathcal{P}_X$ sur E : $\mathcal{P}_X = \{ E_i \}_{1 \leq i \leq k}$, où $\forall i$, $E_i = \{ e \in E \ / \ X(e) = x_i \}$, et on note $n_i = \mathrm{Card}(E_i)$.

Le choix d'un élément e de E peut se faire en choisissant d'abord une partie E_i de E, puis un élément e de E_i.

La réduction du degré d'incertitude sur un élément de E lorsqu'on sait qu'il est dans E_i, (ce qui est la quantité d'information apportée par la classe E_i), vaut : $\mathcal{J}_E(E_i) = \mathcal{J}(E) - \mathcal{J}(E_i) = \log n - \log n_i = \log \frac{1}{p_i}$, [2]

où $p_i = \frac{n_i}{n}$ est la probabilité d'obtenir la modalité x_i si l'on fait un tirage équiprobable sur E.

---

2) Remarque : on utilisera souvent les propriétés des logarithmes:

(L) $\log(ab) = \log a + \log b$ et $\log(a/b) = \log a - \log b$.

On rappelle que $\lim_{x \to 0} (x \log x) = 0$, d'où la convention: $0 \log 0 = 0$.

La quantité d'information apportée sur E par le caractère X est la réduction moyenne du degré d'incertitude : $\mathcal{J}(X) = \sum_i p_i \, \mathcal{J}_E(E_i)$, ou encore:

$$\mathcal{J}(X) = \sum_{i=1}^{k} p_i \log \frac{1}{p_i}$$ **(formule de Shannon, 1948)**

On dit que $\mathcal{J}(X)$ est l'entropie du caractère X ou de la partition $\mathcal{P}_X$ associée à X. On peut montrer que $\mathcal{J}(X)$ est aussi le nombre minimum moyen de questions dichotomiques que l'on doit poser pour déterminer une modalité du caractère X.

On remarquera que $\mathcal{J}(X)$ ne dépend que de la distribution de probabilité $(p_i)_i$ et non de l'ensemble E, on pourra donc généraliser la formule de Shannon donnant $\mathcal{J}(X)$ pour un caractère X défini sur un ensemble quelconque.

3 . PROPRIETES DE L'ENTROPIE

Lemme : Soient $(p_i)_{1\le i\le k}$ et $(q_i)_{1\le i\le k}$ 2 distributions de probabilité, c'est-à-dire 2 systèmes de nombres réels vérifiant :

$\forall i,\ 1 \le i \le k,\ p_i \ge 0,\ q_i \ge 0$ et $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} q_i = 1$; alors :

$\sum_{i=1}^{k} p_i \log \frac{1}{p_i} \le \sum_{i=1}^{k} p_i \log \frac{1}{q_i}$, avec égalité si et seulement si (=ssi) $\forall i,\ p_i = q_i$.

En effet, $\forall\ x > 0$, $\mathrm{Ln}\ x \le x - 1$ avec égalité ssi $x = 1$, d'où :

$\forall i,\ 1\le i\le k,\ \mathrm{Ln}\ \frac{q_i}{p_i} \le \frac{q_i}{p_i} - 1$, ce qui est $p_i\ \mathrm{Ln}\ \frac{q_i}{p_i} - q_i + p_i \le 0$, avec égalité ssi $p_i = q_i$, d'où par sommation, $\sum_{i=1}^{k} p_i \mathrm{Ln}\ \frac{q_i}{p_i} \le 0$, avec égalité ssi $\forall i,\ p_i = q_i$ puis le passage à "log" en divisant par Ln 2.

Propriété 1 : Pour tout caractère X de distribution $(p_i)_{1\le i\le k}$, on a:

$\mathcal{J}(X) \le \log k$ avec égalité ssi $\forall i,\ p_i = \frac{1}{k}$, c'est-à-dire qu'un caractère ayant k modalités apporte le maximum d'information sur E lorsque les modalités sont équiprobables.

On a vu que, à un caractère X sur E ayant un nombre fini de modalités, on peut associer une partition $\mathcal{P}_X$ de E dont les classes sont les parties $\{e \in E / X(e) = x\}$, notées plus simplement $\{X = x\}$.

On dira qu'*un caractère* Y *est déduit d'un caractère* X s'il existe une surjection f de l'ensemble M_X des modalités de X sur l'ensemble M_Y des modalités de Y, ou encore que la partition $\mathcal{P}_X$ est *plus fine que* $\mathcal{P}_Y$, ce qu'on peut noter : $Y = f \circ X \iff Y \propto X \iff \mathcal{P}_Y \propto \mathcal{P}_X$. Dans ce cas, on connait la réponse à Y dès qu'on obtient une réponse pour X.

Le tableau croisé des caractères X et Y, avec X en lignes et Y en colonnes, est alors tel que chaque ligne n'a qu'une seule valeur non nulle. En particulier, si X et Y ont le même nombre de modalités, il y a bijection de M_X sur M_Y : X et Y représentent le même caractère et correspondent à deux formulations différentes de la même question.

Propriété 2 : Si le caractère Y est déduit du caractère X, $\mathcal{J}(Y) \leq \mathcal{J}(X)$.

On notera P la probabilité uniforme sur E.

On procède par induction sur les modalités de Y : pour chaque modalité y_j de Y, soient x_i, $i \in I_j$, les modalités correspondantes de X ;

$$\{Y=y_j\} = \bigcup_{i \in I_j} \{X=x_i\} \text{ , d'où } P\{Y=y_j\} = \sum_{i \in I} P\{X=x_i\} \quad (*)$$

$$\forall\, j,\ \forall\, i \in I_j,\ \frac{P\{X=x_i\}}{P\{Y=y_j\}} \leq 1 \text{ , d'où } \log \frac{P\{X=x_i\}}{P\{Y=y_j\}} \leq 0 \text{ et}$$

$$\sum_{i \in I} P\{X=x_i\} \log \frac{P\{X=x_i\}}{P\{Y=y_j\}} \leq 0 \text{ ; alors en tenant compte de (L) et de (*):}$$

$$P\{Y=y_j\} \log \frac{1}{P\{Y=y_j\}} \leq \sum_{i \in I_j} P\{X=x_i\} \log \frac{1}{P\{X=x_i\}} .$$

On obtient le résultat voulu par sommation sur j.

4 . ENTROPIE JOINTE.

Soient X et Y deux caractères sur E, de modalités $(x_i)_{1 \leq i \leq k}$ et $(y_j)_{1 \leq j \leq l}$. La partition associée au couple (X,Y) est :

$$\mathcal{P}_{(X,Y)} = \mathcal{P}_X \vee \mathcal{P}_Y = \{ \{X = x_i, Y = y_j\} / \ 1 \leq i \leq k,\ 1 \leq j \leq l \}$$

On note $p_{i.} = P\{X = x_i\}$, $p_{.j} = P\{Y = y_j\}$, $p_{ij} = P\{X = x_i, Y = y_j\}$.

L'entropie du couple (X, Y) est appelée *entropie jointe* de X et Y et vaut :

$$\mathcal{J}(X * Y) = \sum_{i,j} p_{ij} \log \frac{1}{p_{ij}} .$$

Propriété 3 : $\mathcal{J}(X * Y) \leq \mathcal{J}(X) + \mathcal{J}(Y)$ avec égalité ssi X et Y sont indépendants.

On applique le lemme aux familles $\{p_{ij}\}$ et $\{q_{ij} = p_{i.}\ p_{.j}\}$.

Propriété 4 : $\mathcal{J}(X * Y) \geq \max\left[\mathcal{J}(X), \mathcal{J}(Y)\right]$ avec égalité ssi un des caractères se déduit de l'autre.

5 . ENTROPIE CONDITIONNELLE ET GAIN D'INFORMATION.

On considère une partie E' ayant n' éléments d'un ensemble E ayant n éléments. Un caractère X sur E partage E en classes $E_i = \{X = x_i\}$ ayant n_i éléments et E' en classes $E'_i = E_i \cap E'$ ayant n'_i éléments, $1 \le i \le k$.

On note X' la restriction de X à E', $p_i = \frac{n_i}{n}$, $p'_i = \frac{n'_i}{n'}$.

La quantité d'information apportée par X pour déterminer un élément qu'on sait être dans E' est $\mathcal{I}(X|E') = \mathcal{I}(X')$, qu'on appellera *entropie conditionnelle* de X sur E'.

La quantité d'information apportée par X pour déterminer un élément de E' dont on ignore qu'il est dans E' est la réduction moyenne du degré d'incertitude des classes E_i pondérées par leurs probabilités sur E' :

$$\mathcal{I}(X,E') = \sum_{i=1}^{k} \frac{n'_i}{n'} \cdot \mathcal{I}_E(E_i) = \sum_{i=1}^{k} p'_i \log \frac{1}{p_i} .$$

Le gain d'information apporté par la connaissance que l'élément est dans E' est: $\Delta\mathcal{I}(X,E') = \mathcal{I}(X,E') - \mathcal{I}(X|E') = \sum_{i=1}^{k} p'_i \log \frac{p'_i}{p_i}$: information de Kullback

D'après le lemme, $\Delta\mathcal{I}(X,E') \ge 0$: c'est bien un gain, et ce gain est nul ssi $\forall i$, $p'_i = p_i$.

Si X et Y sont deux caractères sur E, on notera $p(x_i|y_j)$ la probabilité de $\{X = x_i\}$ conditionnelle à $\{Y = y_j\}$, alors :

$$\mathcal{I}\left[X|\{Y = y_j\}\right] = \sum_{i=1}^{k} p(x_i|y_j) \log \frac{1}{p(x_i|y_j)}$$

L'*entropie* de X *conditionnelle* à Y est la moyenne des entropies de X conditionnelles à chaque modalité de Y :

$$\mathcal{I}(X|Y) = \sum_{j=1}^{l} p_{.j}\, \mathcal{I}\left[X|\{Y = y_j\}\right] = \sum_{i,j} p_{ij} \log \frac{1}{p(x_i|y_j)} , \quad \text{car } p(x_i|y_j) = \frac{p_{ij}}{p_{.j}}$$

Propriétés 5 de l'entropie conditionnelle

$\mathcal{I}(X * Y) = \mathcal{I}(X|Y) + \mathcal{I}(Y) = \mathcal{I}(X|Y) + \mathcal{I}(X)$

$\mathcal{I}(X|Y) \le \mathcal{I}(X)$ avec égalité ssi X et Y sont indépendants.

La 1ère propriété résulte de: $\forall i,j$, $p_{ij} = p(x_i|y_j)\, p_{.j} = p(y_j|x_i)\, p_{i.}$, et de (L), d'où les égalités proposées par sommation sur i et j.

La 2ème propriété résulte de la propriété 3

6 . INFORMATION MUTUELLE

Soient X et Y 2 caractères sur E. Le gain d'information apporté sur X par la modalité y_j de Y est :

$$\Delta \mathcal{J}\left[X \mid \{Y = y_j\}\right] = \sum_i p(x_i \mid y_j) \log \frac{p(x_i \mid y_j)}{p_{i.}}$$

Le gain d'information apporté sur X par le caractère Y est la moyenne de ces valeurs : $\Delta \mathcal{J}(X \mid Y) = \sum_j p_{.j} \, \Delta \mathcal{J}\left[X \mid \{Y = y_j\}\right] = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{i.} \, p_{.j}}$

On voit qu'il y a symétrie en X et Y : $\Delta \mathcal{J}(X \mid Y) = \Delta \mathcal{J}(Y \mid X)$

Cette valeur commune notée $\mathcal{J}(X,Y)$ sera appelée *information mutuelle* de X et Y et vaut : $\mathcal{J}(X,Y) = \mathcal{J}(X) - \mathcal{J}(X \mid Y) = \mathcal{J}(X) + \mathcal{J}(Y) - \mathcal{J}(X*Y)$

On peut interpréter $\mathcal{J}(X,Y)$ comme le gain d'information obtenu quand on passe de la connaissance des distributions marginales $\{p_{i.}\}$ et $\{p_{.j}\}$ à celle de la distribution jointe $\{p_{ij}\}$.

Propriétés 6

a) $\mathcal{J}(X,Y) \geq 0$ avec égalité ssi X et Y sont indépendants.

b) $\mathcal{J}(X,Y) \leq \min\left[\mathcal{J}(X), \mathcal{J}(Y)\right]$, avec égalité ssi un des caractères se déduit de l'autre.

a) découle de $\mathcal{J}(X,Y) = \mathcal{J}(X) + \mathcal{J}(Y) - \mathcal{J}(X*Y)$ et de la propriété 3
b) découle de la propriété 4.

L'information mutuelle $\mathcal{J}(X,Y)$ peut être considérée comme un indicateur de dépendance : pour des caractères X et Y dont les distributions marginales sont fixées, donc aussi les entropies $\mathcal{J}(X)$ et $\mathcal{J}(Y)$, $\mathcal{J}(X,Y)$ est d'autant grand et se rapproche de min $\left[\mathcal{J}(X), \mathcal{J}(Y)\right]$ que les caractères X et Y sont dépendants.

7 . DISTANCE ENTRE CARACTERES.

On définit une distance d sur l'ensemble des caractères sur E, ou ce qui est équivalent sur l'ensemble des partitions de E par :

$d(X,Y) = 2\,\mathcal{J}(X,Y) - \mathcal{J}(X) - \mathcal{J}(Y)$

Propriétés

$d(X,Y) \leq \mathcal{J}(X*Y)$ avec égalité ssi X et Y sont indépendants

$d(X,Y) \geq |\mathcal{J}(X) - \mathcal{J}(Y)|$, avec égalité ssi un des caractères se déduit de l'autre.

8 . LIEN ENTRE 2 CARACTERES.

On pose $Z_{ij} = \dfrac{p_{ij} - p_{i.}p_{.j}}{p_{i.}p_{.j}}$, d'où $\dfrac{p_{ij}}{p_{i.}p_{.j}} = 1 + Z_{ij}$ et $\sum_{i,j} p_{i.}p_{.j}Z_{ij} = 0$

alors : ($\forall$i,j, Z_{ij} est petit) $\Leftrightarrow$ la distribution jointe $(p_{ij})_{i,j}$ est "proche" du produit des distributions marginales $(p_{i.}p_{.j})_{i,j}$

En faisant un développement limité des logarithmes $\mathrm{Ln}\,(1+Z_{ij})$ et en tenant compte de $\log u = \dfrac{\mathrm{Ln}\ u}{\mathrm{Ln}\ 2}$ et de $p_{ij} = p_{i.}p_{.j}\,(1 + Z_{ij})$, on obtient :

$$\mathcal{I}(X, Y) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{i.}p_{.j}} = \frac{1}{2\ \mathrm{Ln}\ 2} \sum_{i,j} p_{i.}p_{.j}\ (Z_{ij}^2 + \varepsilon Z_{ij}^2)$$

On appellera *lien* de X et Y le premier terme de ce développement, au coefficient près :

$$\mathrm{Lien}\,(X,Y) = \sum_{i,j} p_{i.}p_{.j}Z_{ij}^2 = \sum_{i,j} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

Le lien entre X et Y est donc le coefficient φ^2 de Pearson.

On a ainsi la relation suivante entre le lien de X et Y et la valeur du χ^2 calculé sur le tableau des $\{p_{ij}\}$:

$\mathrm{Lien}\,(X, Y) = \dfrac{\chi^2\,(X,Y)}{n}$, n étant le nombre d'éléments de l'ensemble.

C'est aussi le carré de la distance entre la distribution $\{p_{ij}\}_{ij}$ et la distribution produit $\{p_{i.}p_{.j}\}_{i,j}$, pour la métrique du χ^2 centrée sur la distribution marginale $\{p_{.j}\}_j$.

Propriétés 8 :

- Lien (X,Y) = 0 ssi les caractères X et Y sont indépendants.
- Lien (X,Y) $\leq$ min (k - 1, l - 1) , avec égalité ssi un des caractères se déduit de l'autre.

SOURCES :

R. Ash : Information theory, Interscience pub, 1965

A. Rényi : Calcul des probabilités, Dunod, 1966

M. Volle : Analyse des données, Chap. III, Economica, 1985

ANNEXE D

HOMME MOYEN - HOMME TYPIQUE

1 - Le choix collectif

Devant la difficulté d'appréhender dans sa diversité un ensemble d'objets ou d'individus dissemblables, l'idée commune est de synthétiser l'information, c'est-à-dire de dégager la ou les caractéristiques principales des objets ou individus constituant cet ensemble. En particulier, l'idée de choisir un représentant est très ancienne, mais la nécessité et la possibilité de déterminer un tel élément ne sont apparues que progressivement.

Pour choisir un individu comme représentant d'un groupe, la procédure de vote est connue et appliquée depuis longtemps. Cependant, dans le cas du choix de plusieurs représentants pour ce groupe, diverses méthodes sont possibles et des difficultés d'application apparaissent parfois. Cela a été mis en évidence dès le 18e siècle par Borda et Condorcet (voir [2], [4], [8], [12] ([1])).

La question qui est posée est alors: "*peut-on obtenir un choix collectif à partir de préférences individuelles sur les candidats ?*", problème qu'on appelle *agrégation des préférences.*

Borda propose de comptabiliser pour chaque candidat le rang qui lui est attribué par chaque votant. Le classement collectif qui en résulte est obtenu à partir de la somme des rangs.

Condorcet [4] analyse la méthode de Borda et indique qu'elle est très simple d'application et donne toujours un résultat ; mais elle donne quelquefois un classement non conforme au voeu de la pluralité: c'est l'*effet Borda.*

Condorcet propose de décomposer tout problème de vote en questions dichotomiques (oui/non) et d'adopter la règle majoritaire pour chacune de ces questions. Ainsi, effectuer un classement sur les candidats A, B, C revient à se prononcer sur les trois questions: "A est-il préféré à B ?", "A à C ?", "B à C ?", en supposant que chaque votant fait son choix de façon non contradictoire.

([1]) Les nombres entre crochets renvoient à la bibliographie placée en fin d'annexe

Cependant, Condorcet remarque que sa méthode peut aboutir à des propositions contradictoires. On peut avoir les propositions adoptées à la majorité: "A avant B", "B avant C", "C avant A" qui sont globalement contradictoires. C'est l'*effet Condorcet* selon la dénomination de Guilbaud [8].

Un des apports importants de Condorcet sur le plan théorique, est la méthode de décomposition d'une proposition composée en plusieurs propositions simples dichotomiques: cette méthode est très utilisée actuellement sous le nom de réduction à un système disjonctif complet.

2 . L'homme moyen

La notion d'"homme moyen" a été introduite par Buffon [3] d'un point de vue philosophique: "... *d'après les tables de mortalité qui ne représentent jamais que l'homme moyen, c'est-à-dire les hommes en général, bien portants ou malades, sains ou infirmes, vigoureux ou faibles, ...*". De façon intuitive, les tables de mortalité -étant calculées sans tenir compte d'autres caractéristiques- représentaient un homme mythique puisqu'elles s'appliquaient à tous les hommes.

En 1835, le statisticien belge Quételet [13] donne sa définition d'un homme moyen: sur chaque individu d'une population ou d'un groupe choisi, on mesure un certain nombre de caractéristiques quantitatives. On considère alors la moyenne arithmétique des valeurs observées et l'homme moyen serait celui qui, pour chacune des caractéristiques, prendrait cette valeur moyenne.

Très vite, cette conception fut critiquée et Cournot en 1843 écrit [5, pp.213-214]: "*on se propose de définir et de déterminer l'homme moyen par un système de moyennes tirées de la mesure de la taille, du poids, des forces, etc., sur des individus en grand nombre. L'homme moyen ainsi défini, bien loin d'être en quelque sorte le type de l'espèce, serait tout simplement un homme impossible, ou du moins rien n'autorise jusqu'ici à le concevoir comme possible.*"

Cournot justifie cette remarque de la façon suivante: "*Si, par exemple, un triangle est assujetti à rester rectangle pendant que ses côtés varient, il y aura une valeur moyenne pour chacun des trois côtés ; mais ces trois moyennes, prises ensemble, ne conviendront pas à un triangle rectangle, ou ne satisferont pas à cette condition, bien connue, que le carré de l'hypoténuse égale la somme des carrés faits sur les deux côtés de l'angle droit.*"

Bertrand [1] reprend ces critiques et donne l'exemple d'une "sphère moyenne" ayant comme rayon, surface, volume la moyenne des éléments correspondants d'un groupe de sphères et conclut: "... *aucune concession n'est possible, nulle sphère n'est difforme. Un homme malheureusement peut l'être et M. Quételet en profite.*"

On peut néanmoins noter que Bertrand, par ailleurs si fin, a rejetté purement et simplement cet essai de représentation, sans chercher à apporter de solution. Il avait pourtant indiqué un exemple de problème mal posé: calculer la probabilité d'un certain événement en prenant au hasard une corde d'un cercle et obtenait trois valeurs différentes pour cet événement suivant la façon de procéder au choix de la corde.

Ces différentes critiques ont jeté le discrédit, aux yeux des scientifiques, sur cette notion d'homme moyen jusqu'à une date récente où le problème a été repris d'une autre manière.

Comme cela avait été noté par Cournot, la détermination d'une valeur typique pour chaque caractéristique prise séparément ne permet pas d'obtenir une valeur typique globale car les relations préalables pouvant exister entre les variables ne sont pas prises en compte. C'est la raison en particulier des effets Borda et Condorcet où l'application de la règle majoritaire est faite en dehors de la relation d'ordre.

La procédure proposée par Quételet n'est applicable que si les individus, caractérisés par leurs valeurs prises pour chacune des variables, est représentable par un point d'un espace vectoriel. Ceci n'était évidemment pas le cas dans l'exemple considéré par Quételet, ni dans ceux donnés par Cournot et Bertrand.

Cependant, malgré ces critiques -émises depuis près d'un siècle et demi-, une procédure analogue à celle de Quételet est employée couramment dans la présentation des sondages faites par les médias.

Pour des questions nominales, on présente comme typique (ou normal?) un individu qui aurait un comportement ou une opinion conformes à ceux de la majorité pour chacune des questions (cf. chap.5).

Traitons cela sur un exemple ; on pose deux questions A et B à un groupe d'individus et on obtient les résultats croisés suivants:

		réponses à la question A			
		a_1	a_2	a_3	
réponses à la question B	b_1	25	4	23	52
	b_2	4	30	0	34
	b_3	13	0	1	14
		42	34	24	100

Ce tableau croisé peut correspondre à la répartition d'un échantillon fictif de taille 100 ou à une répartition en pourcentages.

Pour la question A, la réponse a_1 est majoritaire et pour la question B c'est b_1. Cependant, pour les questions A et B considérés simultanément, la réponse majoritaire est (a_2,b_2) avec un pourcentage de 30%, et non pas (a_1,b_1) qui n'obtient que 25%.

Le fait de pouvoir suggérer que (a_1,b_1) doit être le comportement majoritaire, et donc normal, n'est rendu possible et acceptable que parce que les tris croisés d'ordre 2, 3... ne sont pas publiés et peut-être même pas effectués!

On ne peut obtenir une valeur typique globale à partir des valeurs typiques de chaque variable que dans certains cas ; en particulier, c'est vrai lorsque les variables sont indépendantes.

3 - La moyenne et ses extensions.

Depuis l'invention de l'espérance mathématique par Huygens en 1657 (voir la note historique de [9]), un apport de première importance fut la liaison faite par Kolmogorov en 1933 entre le calcul des probabilités et la théorie de la mesure et de l'intégration. Cela a rendu possible l'application des travaux d'analyse générale au calcul des probabilités.

Dans un colloque consacré à la théorie des probabilités en 1938, Fréchet lie l'évolution des recherches sur le calcul des probabilités à celle d'autres parties des mathématiques. Dès le début du 20e siècle, l'idée directrice en topologie, en théorie des groupes... est d'étendre les notions à des espaces plus généraux et d'obtenir les propriétés sous des conditions plus larges: "*on envisage d'abord des éléments de nature quelconque et on ne fait intervenir les caractéristiques propres à chaque nature d'éléments qu'au fur et à mesure des besoins et -quand c'est possible- seulement au moment où cela devient indispensable.*"

Dans le cas de variables aléatoires à valeurs réelles, la notion de moyenne a été généralisée en moyenne d'ordre r et z-moyenne (cf. [7b], [11]).

Dès la fin du 19e siècle, des études multidimensionnelles étaient faites (Bravais en France, Galton, Pearson, Fisher en Angleterre), mais les différentes variables étaient considérées comme définissant les coordonnées de points dans un espace vectoriel.

L'extension de la moyenne, liée à celle de la notion d'intégrale, a donc été faite pour des variables aléatoires à valeurs dans un espace vectoriel de dimension finie, puis infinie sous certaines conditions (*): espace de Hilbert (muni d'un produit scalaire et complet) ou espace de Banach-Wiener (espace normé complet)...

Ces diverses extensions prennent en compte principalement la propriété de linéarité venant de la structure vectorielle et ne permettent pas de généraliser la moyenne d'ordre r pour $r \neq 1$.

La possibilité de définir un élément typique exempt des critiques portées contre l'*homme moyen* de Quételet va venir d'une autre approche, déjà connue mais non exploitée.

(*) Fréchet, Intégrale définie sur un ensemble abstrait, 1915 et Intégrale abstraite sur un espace abstrait..,Revue Scientifique, 1944. Voir aussi [6]

4 - Valeurs typiques en tant qu'optimum

Laplace note en 1774 [10a] : *"Par le milieu que l'on doit choisir entre plusieurs observations, on peut entendre deux choses qu'il importe également de considérer: la première est l'instant tel qu'il soit également probable que le véritable instant du phénomène tombe avant ou après ; on pourrait appeler cet instant milieu de probabilité* [actuellement médiane], *...la seconde est l'instant tel qu'en le prenant pour milieu, la somme des erreurs à craindre, multipliées par leur probabilité, soit un minimum ; on pourrait l'appeler milieu d'erreur ou milieu astronomique"* [ou encore moyenne arithmétique].

Laplace revient sur la loi des erreurs et écrit en 1781 [10b]:

> Maintenant, on peut entendre une infinité de choses différentes par le *milieu* ou le *résultat moyen* d'un nombre quelconque d'observations, suivant que l'on assujettit ce résultat à telle ou telle condition. Par exemple, on peut exiger que ce milieu soit tel que la somme des erreurs à craindre en *plus* soit égale à la somme des erreurs à craindre en *moins;* on peut exiger que la somme des erreurs à craindre en plus, multipliées par leurs probabilités respectives, soit égale à la somme des erreurs à craindre en moins, multipliées par leurs probabilités respectives. On peut encore assujettir ce milieu à être le point où il est le plus probable que doit tomber le véritable instant du phénomène, comme M. Daniel Bernoulli l'a fait dans les Mémoires cités : en général, on peut imposer une infinité d'autres conditions semblables qui donneront chacune un milieu différent; mais elles ne sont pas toutes arbitraires. Il en est une qui tient à la nature du problème et qui doit servir à fixer le milieu qu'il faut choisir entre plusieurs observations : cette condition est que, en fixant à ce point l'instant du phénomène, l'erreur qui en résulte soit un minimum;

On voit dans ce passage que Laplace envisageait aussi des moyennes d'ordre *r*, au moins pour *r* entier, et d'autre part que pour lui, la recherche d'une valeur typique se ramène à un problème d'optimum (*).

Cette méthode d'obtention va être mise en oeuvre dans la définition de valeurs typiques donnée par Fréchet en 1940 (voir [7b']).

Soit $r>0$; si X est une variable aléatoire numérique ayant un moment d'ordre r - c'est-à-dire que l'espérance mathématique $E(|X|^r)$ est finie-, alors une valeur typique d'ordre r est un nombre t_r qui réalise la condition d'optimum:

$\forall$ y dans R, $\quad E(|X - t_r|^r) \leqslant E(|X - y|^r)$

En particulier, pour $r=1$ on obtient la médiane et pour $r=2$ la moyenne ou espérance mathématique de X.

(*) Cette recherche d'un optimum était aussi pratiquée avec la méthode des moindres carrés (Legendre 1805, Gauss 1812).

Fréchet montre que si r > 0 et E ($|X|^r$) est finie, une telle valeur typique d'ordre r existe, et qu'elle est unique si r > 1. De plus l'existence du moment d'ordre r - 1 suffit pour avoir une valeur typique t_r, quand $r \geq 1$.

La méthode de Fréchet, -définir une valeur typique comme étant une valeur qui est "la plus proche" de la variable aléatoire parmi l'ensemble de toutes les valeurs possibles-, peut se généraliser à un ensemble non numérique (**). Pour cela, il faut pouvoir définir une distance entre éléments de l'ensemble, et il y a alors de nombreuses possibilités.

Dans le cas d'un questionnaire, chaque individu est caractérisé par ses réponses données aux différentes questions, c'est-à-dire par un *patron*. On notera **R** l'ensemble de tous les patrons de réponses possibles à l'enquête.

On peut définir l'écart entre deux individus comme étant égal à la distance entre leurs patrons.

Les questions étant affectées d'un coefficient α_k qui reflète l'importance que leur attribue le statisticien et/ou l'utilisateur, une distance entre les patrons r_i et r_j peut être définie par :

$$d(r_i, r_j) = \sum_k \alpha_k d_{ij}^k ,$$

où $d_{ij}^k = 1$ si, pour la question k, les modalités sont différentes dans les deux patrons et $d_{ij}^k = 0$ sinon.

Cet écart entre individus est une distance euclidienne, avec pondération des questions, entre les lignes correspondantes du tableau disjonctif complet croisant individus et modalités (voir B 8).

Donnons un exemple simple de ceci:

Considérons une enquête comportant trois questions A,B,C ayant comme modalités respectivement (a_1,a_2,a_3), (b_1,b_2,b_3,b_4), (c_1,c_2).

Il y a alors 3x4x2=24 patrons de réponses. Le calcul de la distance entre les patrons $r=(a_1,b_2,c_1)$ et $r'=(a_1,b_3,c_2)$ nous donne :

$$d(r,r') = \alpha_1 d_{11}^1 + \alpha_2 d_{23}^2 + \alpha_3 d_{12}^3 = \alpha_2 + \alpha_3$$

A partir de cette distance entre patrons de réponses, qui permet de connaître l'écart entre deux individus, on peut déterminer, en utilisant la méthode de Fréchet, un patron typique - et par suite un ou plusieurs individus typiques, représentant de l'ensemble des individus- qui est le plus proche de l'ensemble des patrons.

(**) L'article [7a] de Fréchet est d'une lecture très facile.

Propriété statistique

On tire un échantillon aléatoire de taille n avec remise dans la population, sur lequel on observe les patrons de réponses r_1, r_2,... r_n. On construit un patron typique empirique t_n (non nécessairement unique) qui vérifie:

$$\frac{1}{n} \sum_i d^r(r_i, t_n) = \min_{r \in \mathbf{R}} \frac{1}{n} \sum_i d^r(r_i, r)$$

Le problème que l'on peut se poser est le suivant: est-ce qu'en augmentant indéfiniment la taille de l'échantillon, il y a convergence vers une configuration typique de la population? La réponse est oui dans le cas où cette configuration typique est unique. On dira en théorie statistique que la configuration typique empirique est, dans ce cas, un estimateur convergent de la configuration typique de la population.

Cet écart ainsi défini entre individus permet aussi d'effectuer une classification. Un indice d'agrégation entre deux classes peut être l'écart entre représentants de ces classes, mais il en existe beaucoup d'autres.

Bibliographie de l'annexe D

[1] BERTRAND J. : Calcul des probabilités, Gauthier-Villars, Paris, 1889.

[2] BORDA : Mémoires sur les élections au scrutin, Hist. de l'Acad. des Sciences pour 1781 (publié en 1784), Paris

[3] BUFFON G.L. : Essai d'arithmétique morale, Paris, 1777,
repris dans INET J. et ROGER J. : Un autre Buffon, Hermann, Paris, 1977.

[4] CONDORCET J.A. : Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix, Paris, 1785,
dans CONDORCET : sur les élections, Fayard, Paris, 1986.

[5] COURNOT A. : Exposition de la théorie des chances et des probabilités, Hachette, Paris, 1843.

[6] DOSS S. : Sur la moyenne d'un élément aléatoire dans un espace distancié, Bull. Sc. Math., t73, Paris, 1949.

[7] FRECHET M. :
a - Réhabilitation de la notion statistique de l'homme moyen,
conférence faite au Palais de la Découverte, Paris, 1949.
b - Généralités sur les probabilités - Eléments aléatoires,
Gauthier-Villars, Paris, 1e éd. 1937,
b' - 2e éd. 1950.

[8] GUILBAUD G. : Eléments de la théorie des jeux, Dunod, Paris, 1968.

[9] LANNUZEL, MILLE, ORANGE, PICHARD : Moyennes... vous avez dit moyenne?, IREM de ROUEN, 1987.

[10] LAPLACE P.S. : Oeuvres complètes, Gauthier-Villars, Paris, de 1878 à 1886;
a - Mémoire sur la probabilité des causes..., 1774, t.VIII p.44,
b - Mémoire sur les probabilités, 1781, t.IX p.477.

[11] LEVY P. : Calcul des probabilités, Gauthier-Villars, Paris, 1925.

[12] MICHAUD P. : Hommage à Condorcet, étude IBM, Paris, 1985.

[13] QUETELET A. : Essai de physique sociale, ou sur l'homme et le développement de ses facultés, Paris, 1835.

Pour aller plus loin...

BIBLIOGRAPHIE

Sur les sondages:

[1] DEROO M., DUSSAIX A.M. : Pratique et analyse des enquêtes par sondages, Paris, PUF, 1980.

[2] DROESBEKE, FICHET, TASSI (Editeurs) : Les Sondages, Economica 1987.

[3] GOURIEROUX C. : Théorie des sondages, Economica, 1984.

[4] GROSBRAS J.M. : Méthodes statistiques des sondages, Economica, 1987.

[5] LUBCZANSKI J. : Dossier "Les sondages" in Tangente n°4, 1988.

[6] LUBCZANSKI J.: Les sondages, cuisine "maison" ou expérience scientifique? in bulletin APMEP, n°360, sept 87.

Sur les traitements des données:

[7] CAILLIEZ F., PAGES J.P. : Introduction à l'analyse des données, Paris, SMASH, 1976.

[8] CHANDON, PINSON : Analyse typologique. Théories et applications, Masson, 1980.

[9] DEGENNE A. : Techniques ordinales en analyse des données statistique, Hachette, 1972.

[10] DIDAY, LEMAIRE, POUGET, TESTU : Eléments d'analyse des données, Paris, Dunod, 1981.

[11] FLAMENT C. : L'analyse booléenne des questionnaires, Paris, Mouton, 1976.

[12] HERMAN J. : Analyse des données qualitatives, Masson (Méth.+prog), 1986.

[13] LEBART L., MORINEAU A., FENELON J.P. : Traitements des données statistiques, Paris, Dunod, 1981.

[14] LERMAN I.C. : Classification et analyse ordinale des données, Paris, Dunod, 1981.

INDEX

Titre : Les ENQUETES à QUESTIONS NOMINALES

Réflexions et méthodologie pour l'exploitation d'une enquête à questions nominales.

Auteurs : B. LANNUZEL, G. ORANGE, J.F. PICHARD

Résumé :

Après avoir précisé le vocabulaire et les notions statistiques utilisés: tri-à-plat, tri-croisé, tableau de contingence et tableaux associés, on effectue une analyse critique de sondages publiés dans les médias.

On indique diverses mesures d'association entre deux caractères nominaux, déduites du khi-deux et de la théorie de l'information.

On propose une approche d'un traitement global par une typologie des caractères suivant leur rôle sémantique, la détermination d'un individu typique (homme moyen, portrait robot)et une classification.

Annexe historique sur la notion d'homme moyen et valeur typique.

Mots clés :

statistique
analyse des données
enquête
caractère nominal
tri-à-plat, tri-croisé
coefficient de contingence de Cramer, Pearson, Tschuprow
théorie de l'information
distance du khi-deux
homme typique, portrait robot

Editeur : IREM de ROUEN

Directeur de publication: J.F. PICHARD

octobre 1989 ; format A4 ; 87 pages ; prix : 40 F

n° ISBN : 2-86239-019-4 ; dépôt légal : 4è trimestre 1989