Le Sphinx

Plus^2

Logiciel de traitement d'enquêtes et d'analyse de données

Manuel de référence

Le Sphinx Développement 7, rue Blaise Pascal - 74600 Seynod
Tél. : 04 50 69 82 98 Fax : 04 50 69 82 78 Internet : http://lesphinx-developpement.fr

Le Sphinx Développement

7, rue Blaise Pascal
74600 Seynod

Téléphone : 04 50 69 82 98
Fax : 04 50 69 82 78

e-mail : contact@lesphinx-developpement.fr
Internet : http://lesphinx-developpement.fr

Avant propos

Ce manuel accompagne la nouvelle version du Sphinx que vous venez d'acquérir : Le Sphinx 2000. Nous vous en félicitons.

Ce logiciel s'appuie très fortement sur les interfaces et les protocoles de Windows. Vous retrouverez ainsi les habitudes que vous avez déjà acquises par l'expérience de ce système (gestion des fichiers, menus, dialogues, listes déroulantes, raccourcis…). Nous nous sommes efforcés de toujours respecter ces conventions bien établies. C'est pour cela que vous parviendrez très facilement à l'utiliser.

Vous serez également guidé par votre connaissance du travail des enquêtes, des études et de la statistique. Nous utilisons le langage de ces métiers et avons structuré le logiciel par rapport aux grandes étapes d'une étude : questionnaire, saisie, dépouillement, analyse… Notre souci constant est de faciliter votre compréhension devant l'écran. A cette fin, de nombreux commentaires sont affichés pour vous aider à vous situer et à comprendre les menus, boutons de commande, options… Pour les opérations les plus complexes, vous êtes pris en charge par un assistant. Enfin, lorsque le sens d'un article ou d'un bouton vous échappe, il vous suffira d'essayer pour comprendre très vite son utilité.

Pour toutes ces raisons, vous n'aurez aucune difficulté à vous servir de votre logiciel et peut-être pourrez-vous même vous passer du manuel. Mais vous risquez alors de le sous-utiliser ou de vous compliquer inutilement la vie. En effet, toutes les possibilités qu'il offre ne sont pas également visibles ni toujours très compréhensibles au premier abord et vous risquez de passer à côté de fonctions très utiles.

Le premier objectif de ce manuel est de vous faire découvrir tout ce que vous pourrez faire avec le logiciel. A cette fin, il est organisé en doubles pages développant une tâche, une fonction, un résultat que vous pourrez entreprendre ou réaliser.

La partie de droite vous montre le logiciel, ses écrans, ses menus, ses dialogues et les états (tableaux, graphiques) qu'il permet de produire.

La partie de gauche situe ce que fait le logiciel en terme de finalité, d'utilité et de méthodes mises en œuvre. Vous y trouverez également une explication détaillée des modes opératoires.

Si vous êtes déjà utilisateur du Sphinx, vous pourrez rapidement identifier les nombreuses nouveautés et vous retrouverez facilement vos habitudes.

Si vous découvrez ce logiciel, vous comprendrez facilement son organisation et son système de fonctionnement. La visite guidée que nous vous proposons en tout début du manuel vous accompagnera dans vos premiers pas. En vous reportant ensuite au manuel vous apprendrez à vous en servir complètement et deviendrez des experts.

En vous souhaitant un bon travail.

Yves Baulac, Jean Moscarola

Sommaire

Méthodologie

Rappel des principes de l'enquête par questionnaire

Dans la plupart des cas, l'enquête par sondage est une voie de recherche très efficace pour obtenir des informations. C'est l'outil le plus utilisé dans les études qualitatives et quantitatives, dans les enquêtes et les sondages.
Mais pour être fiable et efficace, cet instrument doit être mené avec précaution. Il est donc nécessaire de déterminer les objectifs de l'étude avant de définir sa mise en œuvre. A cet égard, plusieurs étapes sont alors à envisager.

Définition des objectifs

La première phase d'une étude consiste à définir précisément le problème qui doit être traité. Afin de tirer le meilleur profit des résultats de l'étude, il est indispensable de savoir ce que l'on recherche et d'avoir une idée de l'intérêt de l'étude.
Connaître les données dont on a besoin permet de concevoir un instrument capable de les produire. C'est donc grâce à la détermination préalable des objectifs de l'enquête que des moyens appropriés pourront être définis.

La détermination des échantillons

La réalisation d'un sondage s'effectue sur une partie de la population appelée échantillon. Mais la détermination de l'échantillon doit être calculée afin de fournir des informations et des résultats cohérents avec ceux qui seraient obtenus si toute la population était interrogée. Il faut donc d'abord définir la population mère pour être en mesure d'interroger un échantillon fiable et représentatif.
C'est tout le problème de l'échantillonnage qui consiste à se poser plusieurs questions : Qui interroger ? Combien de personnes ? Quel échantillon retenir et comment le sélectionner (échantillon représentatif ou non, tirage aléatoire ou respect de quotas) ? Dans quelle mesure les résultats de cet échantillon sont-ils fiables ? Etc. Même si les contraintes budgétaires et les techniques employées fournissent des éléments de réponses, la théorie des sondages nous apporte des enseignements précis fondés sur des hypothèses contraignantes concernant la sélection des échantillons. Néanmoins, le savoir-faire et l'expérience guident souvent fortement la détermination de l'échantillon à interroger et les logiciels Sphinx proposent une aide à l'échantillonnage.

Les différentes formes d'enquêtes

Lorsque les objectifs de l'étude, la population mère et l'échantillon sont fixés, il s'agit de définir les moyens mis en œuvre pour l'étude. Si on envisage que l'enquête sera la méthode d'observation et de recueil des informations, on doit alors opter pour une technique d'enquête.
Or celles-ci sont très diversifiées. Parmi les plus répandues, nous retiendrons l'enquête par voie postale, l'entretien téléphonique, le questionnaire sur Minitel, l'enquête sur Internet et l'entretien en face-à-face. Ces derniers peuvent se dérouler à domicile, sur le lieu de travail, dans la rue, à proximité des points de vente...
Le questionnaire est le seul lien, dans les enquêtes par courrier ou par Internet, entre l'enquêteur et la population interrogée. Dans le cas des entretiens téléphoniques ou en face-à-face, la communication entre l'interviewé et l'interviewer passe par le dialogue et le questionnaire devient alors un guide d'entretien ou une grille d'observation où il est possible d'enregistrer les réponses.

Le questionnaire

Le questionnaire est donc destiné à capter, dans la population interrogée, les éléments de réponses aux questions que l'on se pose. Il a alors deux objectifs : provoquer une réaction chez les interviewés et servir de support à l'interviewer qui pourra y enregistrer ses informations de façon complète et précise. La rédaction du questionnaire est à soigner dans ce sens.
Le questionnaire incorpore non seulement les questions à poser mais également les plages de réponses. Il est souvent bon d'indiquer un titre ainsi qu'un commentaire évocateur pour la population interrogée.
C'est un instrument très flexible du fait de la grande variété des questions qui peuvent être posées. L'élaboration d'un bon questionnaire requiert une très grande compétence et peut être intégrée dans des logiciels spécialisés en analyse statistique. C'est le cas du Sphinx qui propose une gamme variée de fonctions destinées à la conception du questionnaire.

Le pré-test

La phase de conception d'une enquête s'achève en général par le test d'une enquête pilote qui permet de valider, sur un nombre restreint de personnes, les choix effectués dans le cadre de l'étude.
Ce test permet de découvrir si le protocole d'étude est réaliste, si le contenu et la forme des questions sont adaptés aux objectifs de l'étude.

C'est aussi souvent l'occasion de découvrir des erreurs grossières et des oublis, ou encore de vérifier la nécessité de chaque question posée et d'écarter éventuellement celles qui ne répondent pas directement aux objectifs de l'étude.
Le test d'une enquête présente donc l'intérêt de rechercher la meilleure adaptation entre les objectifs de l'étude, les moyens alloués et les méthodes choisies.

Le recueil des données

Quoique fastidieuse, cette phase ne présente aucune difficulté particulière, même si selon les techniques d'enquêtes utilisées, les enquêteurs doivent posséder des compétences plus ou moins importantes.
Cette étape nécessite néanmoins une bonne organisation du travail et peut être effectuée dans des logiciels de gestion de données, de traitement d'enquêtes ou d'analyse statistique.
Actuellement, l'amélioration des communications entre ces logiciels permet d'échanger les bases de données très facilement et de reprendre des données existantes comme s'il s'agissait d'informations obtenues par questionnaires. Dans ce domaine, le Sphinx présente une fonction d'importation des données depuis des traitements de textes, des tableurs ou toute base de données externes.

Le dépouillement et l'analyse de données

Une fois les réponses saisies, on s'intéresse aux résultats qui vont ressortir de cette étude. On peut alors distinguer plusieurs niveaux d'analyse : on commencera par le constat des réponses données par les interviewés, c'est-à-dire le dépouillement. Cette phase sera complétée par des calculs ou des tests statistiques et par un approfondissement des analyses pour parvenir aux résultats significatifs de l'enquête, ce qui permettra d'adapter les décisions et les actions aux conclusions de l'étude.
Le dépouillement des résultats donne rapidement un aperçu de l'ensemble des résultats de l'enquête en produisant des tableaux ou graphiques de résultats et des listes de réponses données. Il est d'abord conseillé de prendre connaissance des résultats des variables considérées indépendamment les unes des autres et de procéder ensuite à la mise en relation de plusieurs variables.
La phase d'analyse permet d'effectuer des tests et des calculs sur les résultats extraits du dépouillement. Elle a pour objectif d'analyser les résultats de façon précise et d'aider à l'interprétation et à la décision.
Il est souvent nécessaire, après analyse, de revenir sur la définition initiale d'une ou plusieurs variables pour modifier et enrichir la base initiale de données. A ce niveau, on peut transformer le contenu d'une variable en procédant à des regroupements ou à la suppression de modalités de réponses, mais on peut également créer ou calculer de nouvelles variables.

La présentation des résultats

La présentation des résultats significatifs de l'enquête est parfois une tâche complexe : elle nécessite de prendre connaissance des analyses de données pour ne sélectionner que les plus caractéristiques et les plus synthétiques.
Les résultats retenus pour le rapport d'étude sont également ceux qui sont susceptibles de conduire aux prises de décisions et aux actions. Il s'agit ensuite de les présenter dans un rapport d'étude qui, par sa mise en page, ses commentaires, ses graphiques..., mettra en valeur les résultats significatifs.
Dans cet objectif, le Sphinx propose des fonctions adaptées qui permettent de présenter un rapport organisé selon un plan de dépouillement, de synthétiser et résumer les résultats dans des tableaux construits à cet effet et complétés par des commentaires générés automatiquement par le Sphinx.
Des fonctions de mise en forme des tableaux et graphiques de résultats permettent également de distinguer les résultats les plus significatifs dans l'ensemble des informations ressortant de l'étude réalisée dans le logiciel. Enfin, les nombreuses possibilités d'échange avec les autres logiciels offrent une grande souplesse lors de la réalisation du rapport d'étude.

La communication

Le travail d'enquête et d'étude s'apparente à un travail de communication :

- communication amont pour la transmission du questionnaire, l'exposé des questions, le recueil des réponses ;
- communication avale pour la diffusion des résultats, le travail d'argumentation et d'aide à la décision.

Les nouvelles technologies offrent de nouvelles opportunités de mise en page, de présentation, d'illustration par des couleurs ou des images, d'interactivité dans les échanges avec les répondants ou le destinataire du rapport. Avec l'évolution des logiciels, le chargé d'études peut ainsi de mieux en mieux maîtriser cet aspect important de son travail.

Les différentes formes d'enquêtes

L'observation directe

L'observation directe consiste à mener une observation sans solliciter la participation consciente des personnes observées. Ceci pose bien sûr des questions d'ordre moral : a-t-on le droit de procéder à l'insu de ceux qu'on observe ? Tout dépend de l'usage qui sera fait des informations recueillies.

Il existe aussi des obstacles d'ordre pratique. En effet, beaucoup d'informations sont inaccessibles par cette méthode. D'autre part, les dispositifs concrets permettant d'assurer ce type d'observation (camouflage, glace sans tain, caméra vidéo) sont coûteux et difficiles à mettre en œuvre. Notons cependant les nombreuses possibilités offertes par Internet. L'analyse des traces (origine, pages visitées, temps passé, clic...) laissées par l'internaute est une modalité de l'observation directe.

Entretien en face-à-face

Les protagonistes de l'entretien se font face et peuvent ainsi dialoguer en utilisant toutes les ressources de la communication interpersonnelle. Les circonstances de ce type d'entretien - communication de sujet à sujet - présentent des avantages certains. L'enquêteur sollicite activement le répondant tout en interagissant avec lui pour réguler l'entretien dans sa durée. Des questions peuvent être précisées ou expliquées, l'interprétation des réponses peut être vérifiée, au risque cependant d'influencer ou de biaiser l'observation.

Cette méthode n'est pas exclusive de l'observation directe. L'enquêteur peut, en cours ou à l'issue de l'entretien, noter les caractéristiques du comportement de son interlocuteur. Durée de l'entretien, perception de l'assurance, de la sincérité de l'interlocuteur, présence ou absence de certains indices sur les lieux de l'interview, de comportement a priori définis.

Entretien téléphonique

C'est une autre forme d'entretien. La communication y dispose de moins de ressources. Les protagonistes ne se voient pas, l'enquêteur ignore le cadre dans lequel se trouve le répondant. L'interaction reste possible, mais la bonne compréhension de l'interlocuteur est privée des informations gestuelles. La pression du temps s'exerce différemment.

Enquête par voie postale

Le questionnaire est, dans ce cas, l'unique lien entre l'observateur et la population. Le répondant est seul, libre de répondre ou non, dans l'ordre qui lui convient, sans subir d'autre influence que celle des indications et questions que le questionnaire expose. Il a tout le temps qu'il souhaite pour réfléchir à ses réponses.

L'observateur s'est exprimé une fois pour toutes en élaborant des questions qu'il ne peut plus ni modifier ni expliquer. De même n'a-t-il aucun recours auprès du répondant pour vérifier le sens de ses réponses.

Enquête via Internet

Ce type d'enquête se développe avec l'usage de l'Internet. Le questionnaire est accessible sur un site, le répondant lit les réponses sur son écran et entre directement les réponses. L'avantage de ce procédé est de supprimer en aval la saisie informatique. D'autre part, ce moyen permet de gérer la séquence des questions. Une nouvelle question n'apparaît à l'écran que lorsque la question précédente a reçu une réponse. C'est un avantage par rapport aux enquêtes par courrier dans lesquelles il est impossible de dévoiler progressivement les questions. Cette approche est de plus en plus fréquente pour la consultation des panels. Elle reste encore limitée pour le grand public par le faible taux de connexion des ménages.

Enquête en laboratoire et panel

La situation expérimentale consiste à mettre l'individu dans un contexte contrôlé par l'expérimentateur. Il est possible ainsi, en construisant des plans d'expérience, d'isoler les effets de chacune des actions envisagées.

Avec les panels, on professionnalise l'échantillon en recrutant dans une population considérée, des individus qui acceptent de répondre aux consultations dont ils seront l'objet. Ils sont en général rémunérés et formés au rôle qui est le leur : répondre le plus objectivement possible aux questions qui leur sont périodiquement posées par l'institut qui gère le panel.

	AVANTAGES	INCONVENIENTS
Enquête par observation directe	- Objectivité dans l'observation des faits ou comportements. - Perturbation minimum du fait de l'enquêteur.	- Impossibilité d'observer des opinions ou attitudes. - Difficulté de mise en œuvre pratique (condition de l'observation, formation de l'enquêteur). - Problème déontologique. On observe des gens à leur insu.
Enquête en face à face	- Permet l'observation des attitudes et comportements. - Bon contrôle de l'échantillon sondé : les personnes contactées sont « contraintes » de répondre. - Possibilité de dévoiler progressivement les objectifs de l'enquête. - Souplesse liée à l'enquêteur : adaptation du vocabulaire, interprétation des réponses, précisions apportées. - Possibilité d'entretiens plus longs.	- Coûteux. - L'enquêteur influence le répondant. - Tout dépend de la qualité des enquêteurs, de leur formation à l'enquête, et de leur sérieux sur le terrain.
Enquête téléphonique	- Moins coûteux que face-à-face. - Moins d'influence liée à l'enquêteur.	- Difficulté à poser correctement des questions à réponses assistées. - Impossibilité de passer des questionnaires trop longs.
Enquête par courrier postal ou électronique	- Coût moindre surtout avec Internet. - Le répondant ne subit pas l'influence de l'enquêteur. - Le répondant a le temps de la réflexion, ce qui permet une meilleure approche des questions d'opinion.	- Faible taux de réponse. - Absence de contrôle a priori de l'échantillon. - Forte influence liée au questionnaire et à sa logique.

Un modèle pour concevoir un bon questionnaire

Celui qui rédige un questionnaire peut toujours ramener les questions qu'il envisage à l'un des 4 grands thèmes suivants. Ceux-ci peuvent être étudiés indépendamment les uns des autres, mais la richesse de l'enquête naîtra de la manière dont on est capable de les relier dans un système.

Les grands thèmes d'une enquête

Les quatre grands thèmes suivants peuvent s'appliquer à l'étude de tout type de population. Le 4ème thème ne concerne que les populations humaines.

- **Identité** : qui interroge-t-on? Quels objets observe-t-on?
- **Comportement** : Que font ceux qu'on interroge, comment agissent - ils? Quelles sont les propriétés des objets observés ?
- **Motifs contraintes**: quelles sont les raisons qui guident les comportement, expliquent les actions ? A quelles contraintes, mécanismes sont soumis les objets étudiés ?
- **Opinions et valeurs** : quelle signification les sujets accordent-ils à leur comportements, sur quelles valeurs se fondent leurs motifs d'action ?

Concevoir le questionnaire comme un système

Les thèmes qui structurent le questionnaire peuvent être envisagés comme un système situant les questions les unes par rapport aux autres. Ainsi, l'explication d'un comportement peut être recherchée dans des facteurs d'identité suivant les modèles du déterminisme social ou dans la prise en considération des motifs en référence au modèle de décision rationnelle. Toutes les relations envisageables entre les différents thèmes peuvent faire sens en renvoyant aux grandes théories du domaine étudié.

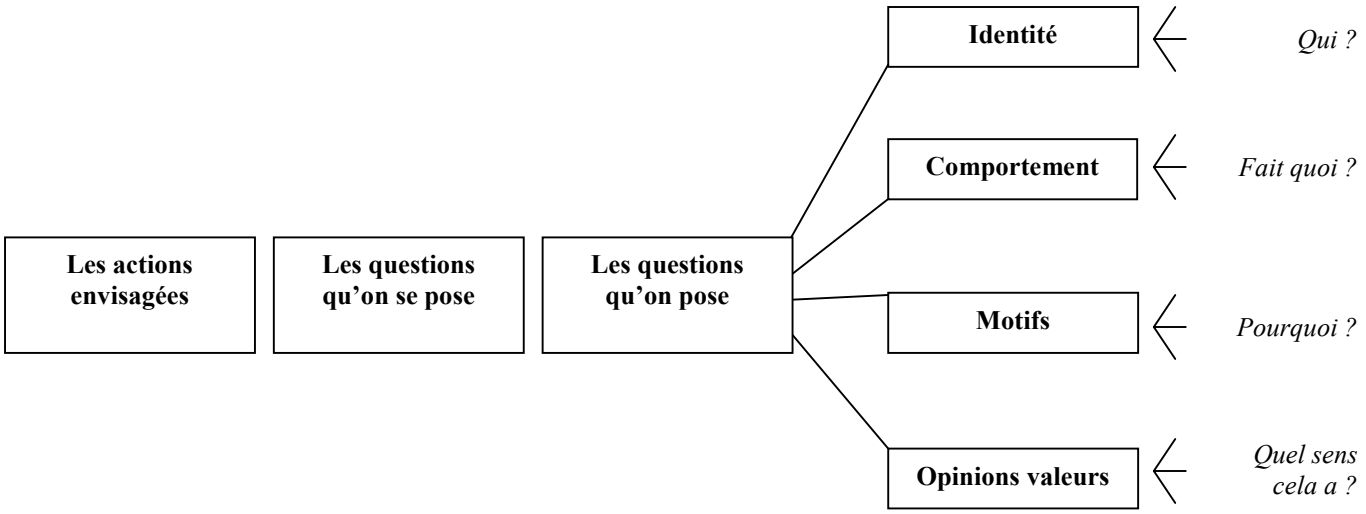
Raisonner ainsi sur les relations entre les questions permet d'affiner le questionnaire en anticipant sur les analyses qui se révéleront utiles au moment de l'analyse des résultats.

Utiliser le modèle identité, comportement, motifs, opinion et valeurs

Quel que soit votre domaine d'étude, il vous sera utile de revenir sur votre questionnaire en l'analysant du point de vue de ce modèle. Tous les thèmes sont-ils abordés, les manques correspondent-ils à un choix délibéré ou à un oubli ?

Quelles sont les relations envisageables entre questions ? A quel type de théorie renvoient-elles ? ....

Le questionnaire comme liste



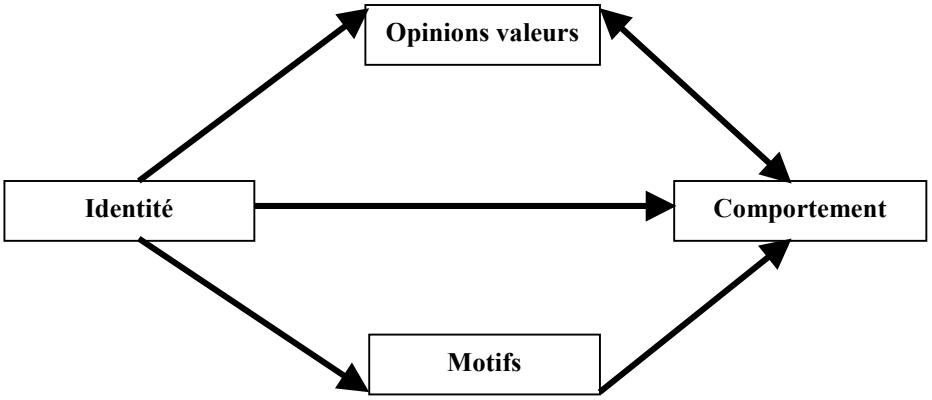
Le questionnaire comme système

Qui fait quoi ?

Quels motifs expliquent les comportements ?

Quel sens accorder au comportement ?

Qui pense quoi ?



Des questions qu'on se pose aux questions qu'on pose

Pièce maîtresse de toute enquête, le questionnaire est à l'articulation du modèle conceptuel, expression des hypothèses et du mode opératoire, instrument d'observation et de recueil de l'information. De la théorie à l'observation, du concept à la variable, le questionnaire organise, par l'intermédiaire des questions posées et des réponses obtenues, la confrontation des idées aux phénomènes du terrain.

Quelle que soit la sophistication des traitements effectués et le sérieux des chiffres, c'est sur le sens des mots que reposent finalement les contributions de l'étude.

La question qu'on se pose

Toute question posée dans un questionnaire renvoie à une ou plusieurs questions que s'est posée celui qui fait l'étude. Ainsi, derrière toute question qu'on pose, il y a une question qu'on se pose : une hypothèse. Celle-ci renvoie à une ou plusieurs théories, connaissances préalables dégagées dans une phase préliminaire d'étude documentaire et de réflexion. La qualité du questionnaire repose sur la clarté de ce travail initial.

La question qu'on pose

Sa fonction est de susciter une réponse, donc de provoquer une réaction. Mais pas n'importe quelle réaction. On recherche en fait la réponse à la question qu'on se pose et on veut la connaître avec une objectivité maximale. Il faut donc perturber le moins possible l'authenticité de ce qu'est celui qu'on interroge, ce qu'il fait, sait, ressent, pense et à tout prix, éviter de provoquer une réponse qui serait influencée par les circonstances de l'interrogation.

Neutralité, objectivité, mais aussi clarté. Que le sens de la question soit le même pour celui qui la pose que pour celui qui l'entend. La qualité des réponses dépend de celle de la compréhension entre questionneur et répondant.

La réponse qu'on enregistre

La finesse de l'observation dépend du procédé d'enregistrement de la réponse. Répondre en choisissant parmi une liste de modalités prédéfinie fait perdre la variété et les nuances que permet l'enregistrement d'une réponse librement formulée. Il en va différemment lorsqu'il s'agit d'une grandeur ou d'un nombre. Il suffit alors d'enregistrer tel quel le chiffre annoncé pour saisir dans l'unité considérée toutes les nuances de la réponse. Dans tous les autres cas, la mesure dépend de l'étalonnage de l'instrument. Choix de l'unité, définition a priori d'un système de codification : dès la conception du questionnaire, il faut imaginer les réponses. Ce travail suppose une connaissance a priori sur les phénomènes abordés. Sans hypothèses, pas d'observation.

Rédiger un bon questionnaire

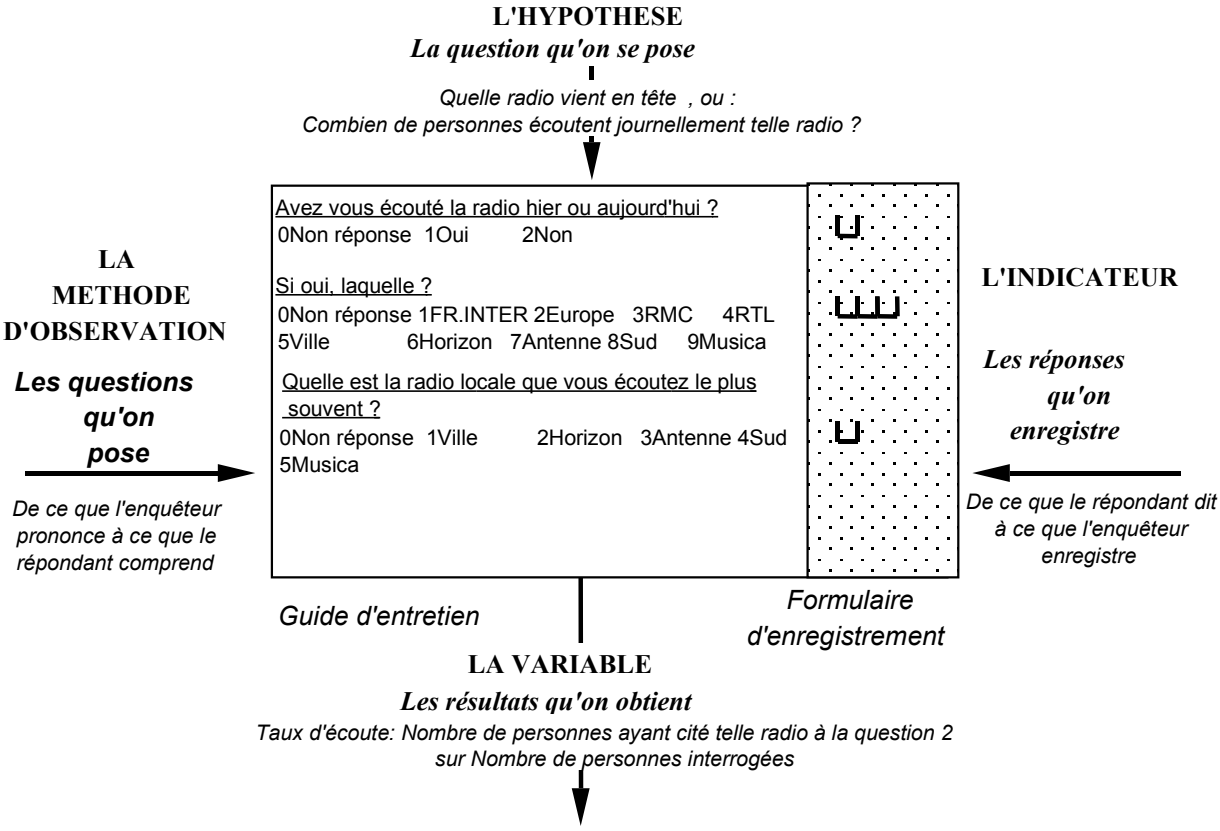
Réaliser un bon questionnaire nécessite de la méthode et beaucoup de métier.

La méthode : respecter les étapes de la démarche :

- Formuler des hypothèses claires, les questions qu'on se pose.
- Traduire ces hypothèses sous forme de questions à poser en tenant compte des caractéristiques des populations interrogées, de leur langage et des circonstances de l'interrogation.
- Tester le questionnaire en l'administrant en situation. Cette phase fait ressortir les problèmes de langage, les difficultés de compréhension, elle permet d'évaluer la durée de l'entretien...
- Dépouiller et analyser les réponses obtenues au cours du test. On pourra ainsi constater que telle question n'est pas vraiment utile, que telle autre manque, que la manière d'enregistrer les réponses n'est pas la plus pratique. On est ainsi conduit à tenir compte, dès la conception du questionnaire, des traitements que l'on souhaite faire.

Le métier : il s'acquiert avec l'expérience. Formuler un bon questionnaire est une tâche difficile qui exige de trouver le meilleur compromis entre les impératifs de la conceptualisation, du terrain et des méthodes statistiques. A cet égard, si vous êtes néophyte, vous gagnerez beaucoup de temps et obtiendrez de bien meilleurs résultats en faisant appel aux conseils d'experts. C'est ce que nous vous proposons au travers du service ***Le Sphinx en direct***. Consultez-nous pour bénéficier de notre expérience.

Le questionnaire est au centre du dispositif d'enquête



La manière de poser les questions

Entretien libre : Question ouverte, réponse spontanée

Une question est à réponse spontanée lorsque aucune indication n'est apportée au répondant sur la manière de formuler sa réponse. Même s'il s'agit d'une question à réponse fermée - dont la réponse sera codée par rapport à des modalités préétablies - la nomenclature préconçue ne lui est pas communiquée. Par exemple, la profession est demandée sans que la classification en CSP ne soit fournie. Le répondant est donc entièrement libre de sa réponse.

Entretien semi ouvert : Question ouverte, réponse fermée

La réponse librement obtenue n'est pas nécessairement enregistrée telle quelle. L'enquêteur peut interpréter la réponse pendant l'entretien, ou à réception des questionnaires expédiés par courrier. Il applique pour ce faire la codification dérivée des modalités de réponses définies avec la question.

Procéder ainsi suppose :

- En face-à-face, que l'enquêteur prenne garde de ne pas dévoiler la liste des modalités qu'il a sous les yeux ;
- Par courrier ou Internet, que les modalités de réponses aux questions qui doivent rester ouvertes, ne figurent pas sur le questionnaire expédié, bien qu'elles aient été préconçues et doivent servir ultérieurement à la codification.

Comme le montre le tableau ci-contre, l'entretien est dit ouvert si la réponse est libre et enregistrée telle qu'elle est donnée. Il est semi-ouvert si la réponse est libre, mais la réponse enregistrée suivant la codification préétablie.

Entretien fermé : Question assistée

La question est fermée ou à réponse assistée si la liste des modalités de réponses est communiquée au répondant. On l'aide à répondre en lui proposant de choisir parmi une liste.

- Dans les enquêtes par courrier ou Internet, les questions sont généralement assistées. Très couramment utilisé, ce procédé simplifie considérablement le traitement. On évite ainsi tout le travail d'interprétation engendré par le système semi-ouvert.
- Dans les enquêtes en face-à-face, l'enquêteur peut énumérer la liste, la faire lire sur son document ou utiliser des panneaux écrits en grosses lettres à cet effet. Cette dernière méthode est de loin la meilleure.

Au téléphone, l'énumération est le seul moyen d'assister la question. Cela ne va pas sans poser quelques difficultés si la liste est trop longue.

Procéder ainsi a l'avantage de la facilité pour ceux qui font l'enquête. Mais cette méthode comporte également des risques.

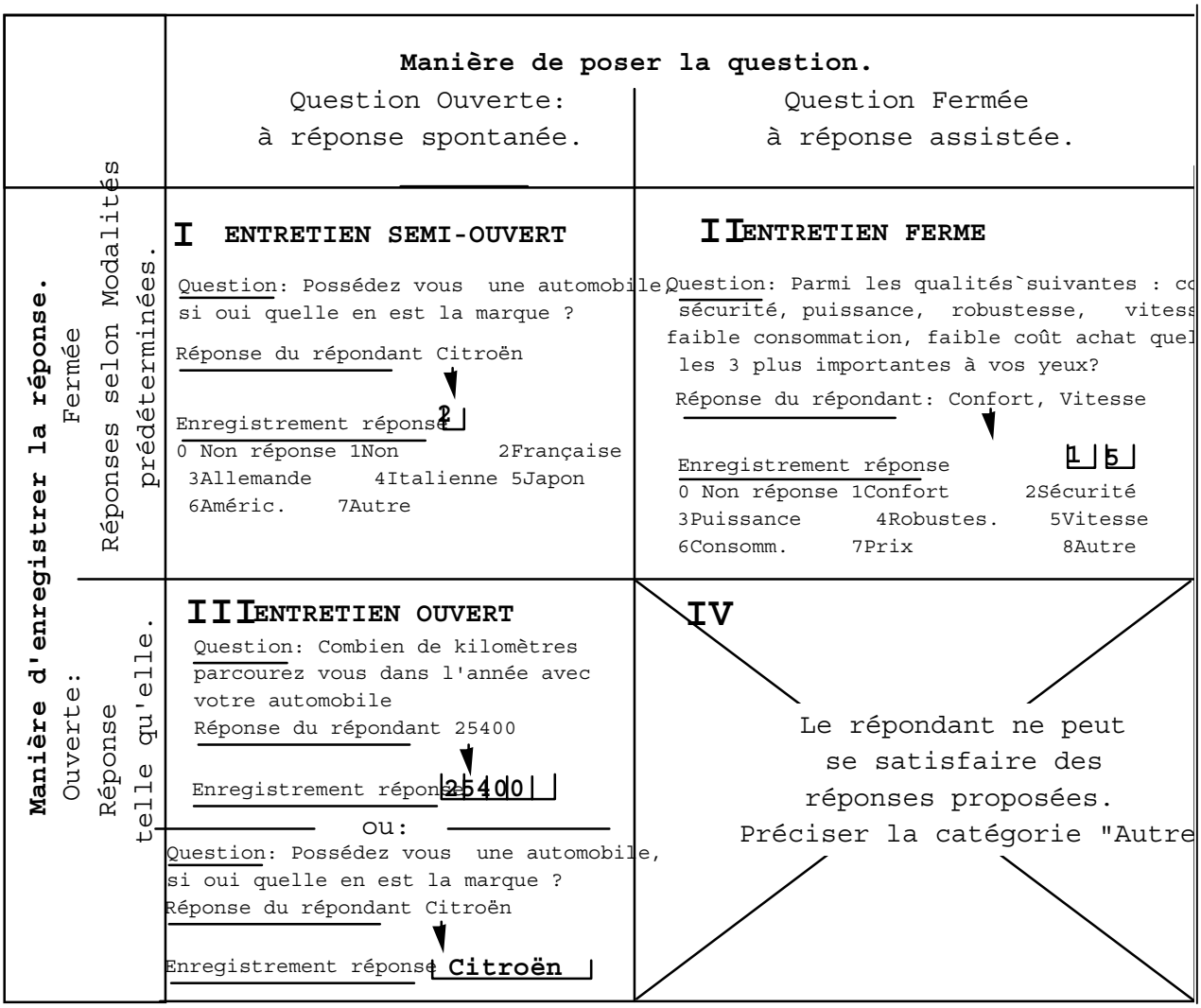
- Le répondant est fortement poussé à répondre. Il faut donc bien préciser que la réponse n'est pas obligatoire et prévoir une rubrique la rubrique "autre précisez"
- L'ordre dans lequel sont énumérées les questions et les modalités de réponses a une influence sur le choix. Lorsque la liste est longue, les dernières modalités sont survolées ou trop rapidement énoncées par l'enquêteur. L'attention se fixe alors sur les premières citations. Si de plus, celles-ci sont des réponses évidentes, le phénomène est amplifié : il faut donc éviter de placer ces modalités en tête de liste.

Pour éviter les autres effets d'ancrage sur le début de la liste, il peut être intéressant de soumettre l'échantillon à des énumérations différentes obtenues par permutation. On neutralise ainsi les éventuelles influences en faisant varier la position des termes.

La manière de poser une question peut en affecter le sens

Suivant que la question est assistée ou non, l'information recueillie n'a pas la même signification. Nous reviendrons ultérieurement sur ce point. Notons pour l'instant que cette décision affecte le fond de l'enquête et ne doit pas être prise à la légère. Son application correcte par les enquêteurs doit donc être soigneusement contrôlée.

Poser les questions et enregistrer les réponses



L'art du questionnaire : la logique de l'entretien

Quelques principes fondamentaux doivent être scrupuleusement respectés.

Introduire le questionnaire

Courrier ou media électronique

Un texte bref situe l'organisme qui réalise l'enquête et donne quelques indications sur les buts poursuivis ainsi que sur la manière de remplir le questionnaire.

Il est important de prévoir une incitation à répondre : explication des buts de l'enquête, proposition de transmettre les résultats, ou cadeau. Tout dépend du budget mais la force de l'incitation peut affecter beaucoup le taux de réponse.

Plus le questionnaire est long, plus faible est l'incitation à répondre. Il doit être clairement rédigé, aéré et occuper un nombre de pages le plus réduit possible.

Les possibilités graphiques et d'animation des médias électroniques peuvent être utilisés comme des incitations à répondre. Mais attention à ne pas surcharger les pages et allonger ainsi le temps de réponse.

Face-à-face

Tout le processus repose sur la qualité des relations que l'enquêteur parvient à établir. Sa mise, son expression, ses attitudes doivent être adaptées au public qu'il interroge. Ces paramètres doivent être adaptés aux circonstances de l'enquête (dans la rue, à domicile...).

Téléphone

Il n'y a pas de différence fondamentale entre la prise de contact en face-à-face et au téléphone mais le téléphone exige une concision et une clarté d'expression encore plus grande. Si le répondant se déclare indisponible, il est plus aisé d'obtenir un rendez-vous téléphonique. Le simple fait de le solliciter permet parfois de faire tomber l'objection.

Respecter les usages de la conversation

Un entretien a sa logique propre. Comme une conversation, il évolue de propos généraux vers des questions plus précises. En face-à-face ou au téléphone, il faut en tenir compte.

Les questions gênantes ou difficiles doivent être reportées en fin d'interview.

Il faut éviter de faire débuter un questionnaire en face-à-face par des questions d'identité. Ce qui se justifie du point de vue de l'analyse risquerait de transformer l'entretien en interrogatoire de police. Il faut prévoir des questions d'introduction ou de transition. Peut être inutiles pour l'analyse, elles ont pour but de rendre l'entretien plus facile.

Enfin, l'ordre des questions doit tenir compte des effets perturbateurs de l'entretien. Les questions à réponse spontanée doivent précéder les questions assistées. Dans le même esprit, il faut révéler le plus tard possible le but précis du questionnaire. En effet, certaines réponses risquent d'être influencées par celui-ci. Même par courrier, il faut respecter cette règle. On veillera également à ce que les questions dont les réponses peuvent être liées figurent sur des pages différentes.

Longueur du questionnaire

Plus un questionnaire est long, plus il est difficile à administrer. Cette contrainte joue en particulier pour les enquêtes par courrier et téléphoniques.

Eviter les questionnaires organigrammes

Il faut éviter de multiplier les questions-renvois. Elles compliquent beaucoup l'entretien et ne se justifient pas toujours. On risque en effet d'imposer au répondant une logique qui n'est pas la sienne et de biaiser ainsi l'observation. Il faut donc se garder de construire un questionnaire comme on conçoit un organigramme et limiter les questions-renvois aux impossibilités strictes de répondre.

L'art du questionnaire : la logique de l'analyse

La difficulté consiste à rédiger un questionnaire qui passe bien et qui permette par la suite les analyses les plus riches possibles.

Ne poser que des questions utiles

C'est une évidence qu'il faut rappeler car la rédaction fait souvent oublier les objectifs initiaux. En confrontant la liste des questions que se pose le demandeur à celle des questions rédigées dans le questionnaire, il faut vérifier que rien n'a été oublié et que tout est nécessaire.

Deux cas peuvent alors se présenter :

- Une question du demandeur de l'étude reste sans réponse. Il faut compléter le questionnaire ou constater qu'il est impossible de répondre sérieusement à l'objectif fixé.
- Une question du questionnaire ne peut être rattachée à aucune des questions du demandeur. Il faut la supprimer ou ajouter aux objectifs telle contribution initialement non prévue. Mais assurons-nous alors que c'est une connaissance susceptible d'affecter l'action du demandeur.

Adapter le questionnaire aux traitements et aux analyses projetés

Nomenclatures et analyse par strate

La qualité des résultats est fonction de la taille de l'échantillon. Ainsi, découper un petit échantillon en strates trop nombreuses conduit à des résultats sans signification.

Il faut par conséquent adapter à la taille de l'échantillon, les modalités des questions définissant les strates. Si N est l'effectif total et si le nombre de modalités dépasse N/30 ces modalités définiront au moins une strate non exploitable. Ainsi, utiliser une nomenclature de C.S.P. en 12 postes en n'interrogeant que 200 personnes nous obligera à regrouper des catégories entre elles pour obtenir des strates significatives.

Type de variables

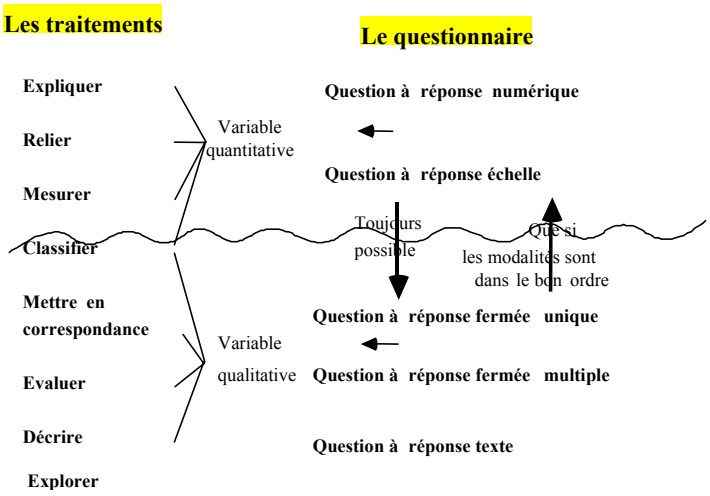
Suivant le type de questions, les réponses obtenues qualifient des états possibles (variables qualitatives), ou mesurent des grandeurs (variables quantitatives).

Les variables quantitatives sont plus riches et autorisent des traitements plus complexes : calculs de moyennes, d'écarts types, corrélations, régressions, analyse en composantes principales. Ces indicateurs ou méthodes permettent des analyses plus fines et plus puissantes que celles autorisées par l'usage des variables qualitatives.

Il faut donc être capable, dès la conception du questionnaire, de définir les traitements envisagés, pour recueillir les informations nécessaires selon le bon format. Ce choix consiste à déterminer le format de la réponse (texte, codée, fermée, échelle, numérique) en fonction des analyses envisagées.

Dès la conception du questionnaire, tenir compte de l'analyse des données

Certains choix effectués au moment de la rédaction du questionnaire conditionnent fortement les possibilités ultérieures d'analyse. Ainsi, s'il est toujours possible de transformer une réponse numérique en variable qualitative, l'inverse n'est pas toujours possible. Si d'autre part, aucune question ouverte n'a été prévue, le questionnaire ne permettra aucune véritable découverte, et sans numérique, aucune mesure n'est facilement réalisable.



Théorie des sondages et estimation statistique

Réaliser un sondage, c'est substituer à l'étude d'une population entière (la population mère), l'observation d'une partie de cette population, l'échantillon. La théorie des sondages permet de :

- définir les conditions selon lesquelles on peut estimer les propriétés de la population totale à partir des observations faites dans l'échantillon.
- qualifier l'estimation en indiquant le degré d'erreur ou de risque qu'elle comporte.

Le sondage aléatoire

L'estimation statistique ne peut être effectuée que sous les conditions d'un sondage aléatoire. Celles-ci impliquent que la population soit de très grande taille par rapport à celle de l'échantillon et que chaque individu de la population ait exactement la même chance de faire partie de l'échantillon. Sous ces conditions, le calcul de probabilité montre qu'on obtient un échantillon dont la composition est voisine de celle de la population mère.

L'estimation statistique

Dans l'hypothèse du sondage aléatoire, on peut calculer, à partir d'un résultat observé dans l'échantillon, l'intervalle dans lequel doit normalement se situer la valeur correspondante dans la population totale. Cette fourchette appelée intervalle de confiance indique ainsi la marge d'imprécision que comporte toute estimation. On a l'habitude de la distinguer du risque d'erreur pris en acceptant cette fourchette comme valide. Plus on souhaite réduire le risque, plus la fourchette sera large et les résultats imprécis, au contraire, on peut désirer afficher des résultats plus précis mais avec un risque d'erreur plus grand. La seule manière d'améliorer la précision sans augmenter le risque d'erreur est d'augmenter la taille de l'échantillon.

Attention: la qualité de l'estimation ne dépend que de la taille de l'échantillon, sous réserve que le tirage est bien aléatoire. Si l'interrogation porte sur une population entière, les résultats sont exacts et il n'y a plus lieu de parler d'estimation.

Si l'exhaustivité n'est pas atteinte, quelle que soit l'importance du taux de réponse, l'estimation n'est possible que si les réponses obtenues sont le fait du hasard. Sa qualité ne dépend que du nombre de répondants.

Les paramètres influençant la qualité d'une estimation

L'intervalle de confiance dépend essentiellement de la taille n de l'échantillon. Par exemple, pour l'estimation d'une proportion p, on le calcule en application de la formule ci-contre. Il est important de constater que l'intervalle de confiance décroît avec la racine carrée de la taille de l'échantillon, ce qui signifie que plus l'échantillon est grand, plus le gain en précision sera faible.

D'autre part, le produit p*(1-p) est de valeur maximum quand p est égal à 0,5 ; ce qui signifie qu'il sera beaucoup plus difficile d'estimer la victoire d'un candidat de deuxième tour au soir des élections (il faudra examiner près de 3000 bulletins) que l'élimination d'un petit candidat au premier tour. Une centaine de bulletins suffisent pour prévoir l'échec d'un candidat rassemblant 10% des suffrages.

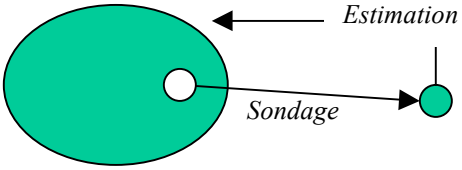
La pratique des sondages

Les conditions théoriques du sondage aléatoire sont très difficiles à réaliser pratiquement. Aucun protocole concret ne permet d'assurer la stricte équivalence des chances d'être interrogé. Même si les numéros de téléphone sont tirés au hasard, la présence ou l'absence au moment de l'appel introduit un biais lié au mode de vie...

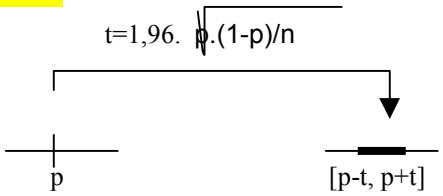
Pour cette raison, il est toujours utile de vérifier la qualité d'un échantillon en contrôlant, sur des caractères connus dans la population totale, que les résultats sur l'échantillon sont conformes. Sinon, on dit que l'échantillon est biaisé. Plutôt que d'avoir à le redresser a posteriori, on peut fixer un plan de sondage par quota pour assurer a priori la proportionnalité de l'échantillon. Les limites de cette méthode tiennent à la connaissance de la population à interroger et aux possibilités pratiques de recueillir des réponses à partir de plans de sondage multi-critères très fins. Interroger tant d'hommes, ouvriers, de plus de 50 ans...

Dans la pratique, on combine souvent la méthode des quotas avec une procédure libre pour trouver les individus correspondant aux quotas prédéfinis.

Tirage aléatoire et estimation statistique



Population mère Echantillon

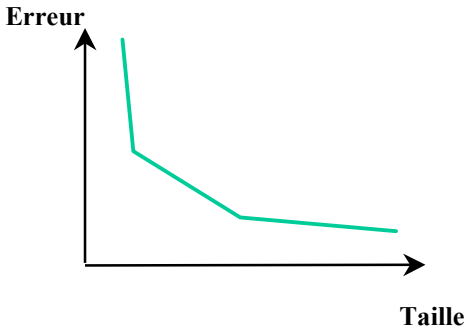


Proportion dans l'échantillon *Intervalle de confiance dans la population*

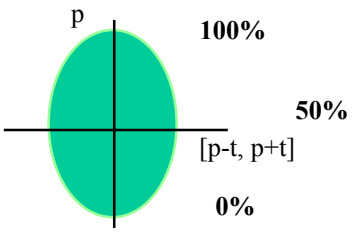
Effet taille de l'échantillon

n=100
n=300
n=1000
n=3000

Les grands échantillons sont plus précis



Effet de l'ordre de grandeur du phénomène



Les phénomènes «grossiers»sont plus faciles à décrire précisément

Définir le bon échantillon

La théorie des sondages nous apprend que l'estimation statistique n'est possible que si l'échantillon est aléatoire et s'il est suffisamment important. L'estimation est interdite si l'échantillon est trop petit : n<30.

Outre les difficultés pratiques du tirage aléatoire, la détermination de la taille de l'échantillon nécessite le choix d'un bon compromis entre la précision attendue et le coût de collecte des données.

Echantillons homogènes

On est dans cette situation lorsque les analyses ne portent que sur l'échantillon total. On cherchera dans ce cas à sélectionner l'échantillon le plus grand possible compte tenu des moyens disponibles. Plus l'échantillon est grand, plus il faudra interroger de personnes supplémentaires pour gagner en précision. A partir d'un certain seuil, les gains en précision ne justifient plus le coût supplémentaire que cela implique.

Tout dépend en fait du type de décision à prendre et de la marge d'incertitude tolérable. Ainsi, on sera beaucoup plus exigeant pour évaluer l'audience d'un média en vue de fixer des tarifs publicitaires que pour une étude de satisfaction. Dans le premier cas, on s'orientera vers de grands échantillons (1000 à 2000 sondés), dans le second, on pourra se contenter d'échantillons plus modestes (200 à 300).

Echantillons hétérogènes

Cette situation correspond au cas où l'on souhaite établir des résultats sur des sous-ensembles de l'échantillon. Si l'échantillon est aléatoire, on obtiendra des effectifs très faibles pour les catégories peu représentées ; ce qui risque d'interdire toute estimation sur les strates correspondantes. Ainsi, pour analyser la strate d'une catégorie représentant 2% de la population totale, il faut un échantillon contenant au moins 30 personnes dans cette catégorie ; ce qui nécessite l'interrogation de 1500 personnes. Mais cet effectif ne se justifie pas pour étudier la population dans son ensemble. Il faudra donc trancher entre renoncer à analyser toutes les strates ou supporter le coût d'interrogation de 1500 personnes.

Une solution de compromis consiste à définir un échantillon stratifié dans lequel on alloue le budget disponible à chacune des strates. Si on dispose d'un budget de 500 personnes et si la population se compose de 5 strates, on interrogera aléatoirement 100 personnes de chaque catégorie. On est ainsi assuré d'avoir une précision convenable pour l'analyse de chaque strate. Mais on ne pourra rien tirer de l'analyse de l'échantillon total dans lequel certaines strates seront sur-représentées et d'autres sous-représentées.

Redressement d'échantillon

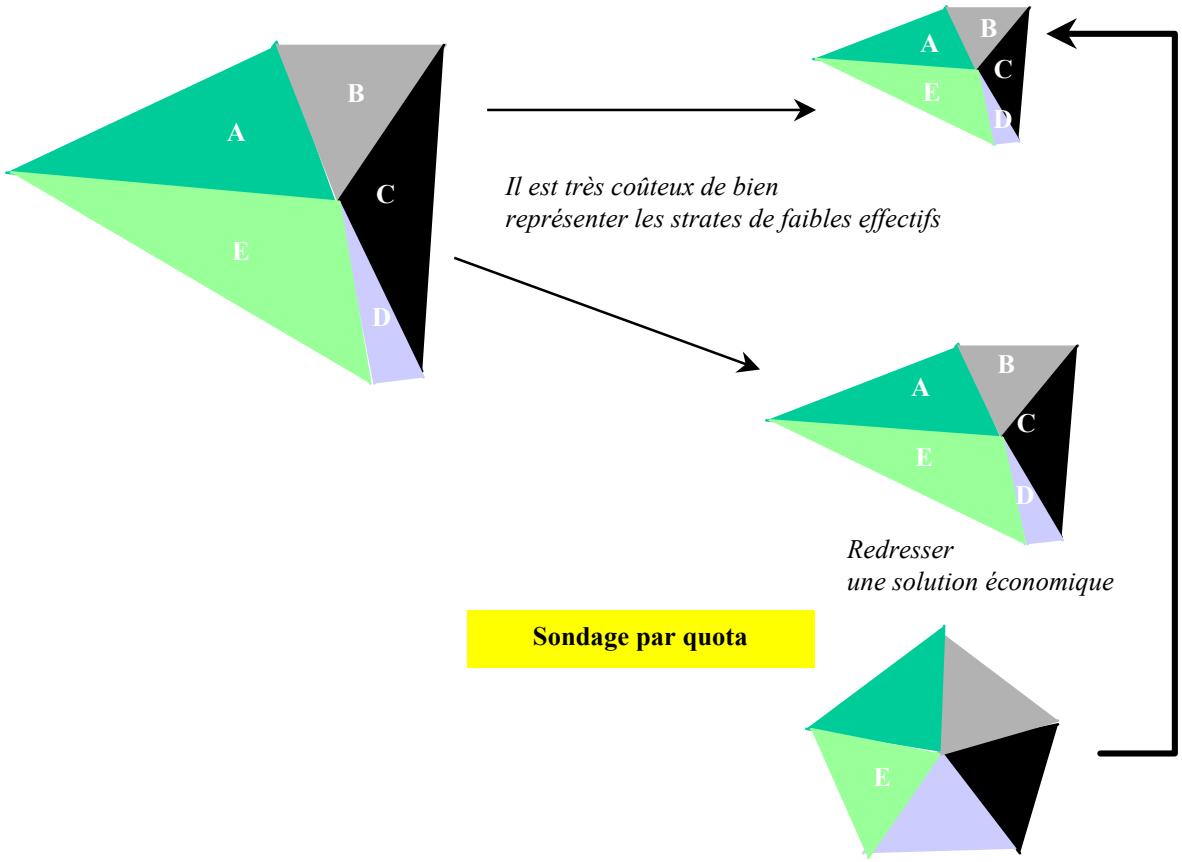
Dans le cas précédent, on redressera l'échantillon total pour composer un nouvel échantillon dans lequel chaque catégorie sera représentée à proportion de son poids dans la population totale.

Procéder ainsi conduit, au niveau de l'analyse, à travailler sur les échantillons propres à chaque strate et sur un échantillon redressé pour analyser la population totale.

Choisir la taille de l'échantillon

Taille de l'échantillon	Taux d'erreur	Intervalle de confiance pour p=50%
200 300	6,93% 5,65%	[43,07% , 56,93%] [44,34% , 55,65%]
2000 3000	2,19% 1,79%	[47,81% , 52,19%] [48,21% , 51,79%]

Sondage aléatoire



Il est très coûteux de bien représenter les strates de faibles effectifs

Redresser une solution économique

Sondage par quota

Les formes de l'analyse de données

Les méthodes de l'analyse de données sont multiples et répondent à des objectifs variés :

- dépouillement visant à restituer les réponses de manière individuelle ou synthétique ;
- transformation des données originales par recodification ou calcul ;
- analyses statistiques visant à décrire, expliquer ou classifier.

Elles peuvent concerner une ou plusieurs variables, un seul individu, tous les individus de l'échantillon ou un sous-ensemble appelé strate.

Le niveau d'analyse

Les dépouillements peuvent se faire à différents niveaux :

- **Au niveau de l'individu**. On s'intéresse aux données de détail en vue d'entreprendre des actions individuelles. Les traitements se ramènent alors à restituer tout ou une partie de l'information propre à chaque observation. Les résultats sont produits sous forme de listes plus ou moins étendues et structurées selon le nombre d'individus décrits. Ces extractions peuvent être effectuées au stade de la saisie (***Consulter un profil d'individus***), ou au stade du dépouillement en produisant des listes.
- **Au niveau de la population entière**. On cherche à synthétiser les informations décrivant chaque individu pour traiter la population comme un tout. Ceci revient à gommer les variations individuelles pour décrire chaque variable par un ou plusieurs indicateurs synthétiques : valeur moyenne pour les variables numériques, fréquence relative des modalités d'une variable nominale.
- **Au niveau d'un sous-ensemble de la population.** On analyse des catégories particulières d'individus pour tenir compte de l'hétérogénéité de la population. En définissant les conditions auxquelles doit répondre le sous ensemble étudié, on construit la strate qui se substitue alors à l'examen de l'échantillon total. La fonction ***Changer de strate*** permet ainsi d'analyser comme un tout n'importe quel sous-ensemble d'individus.

De l'univarié au multivarié

On peut distinguer 3 grandes problématiques auxquelles répondent 3 manières d'aborder l'analyse :

- **L'analyse univariée :** on décrit la population en examinant une seule variable à la fois. C'est la manière la plus simple de restituer l'information et de faire de l'estimation statistique.
- **L'analyse bivariée :** on s'intéresse aux relations existant entre 2 variables à des fins d'explication et/ou de prédiction. Cette approche nécessite la formulation d'une hypothèse que la statistique permettra d'infirmer ou de confirmer.
- **L'analyse multivariée :** on aborde la complexité résultant de la multiplicité des variables.

Dans une approche descriptive, on cherche à réduire le nombre de variables (analyses factorielles) ou à agréger les individus en catégories homogènes (typologies).

Dans une approche explicative, on cherche à intégrer la pluralité des causes et des effets d'interaction (régression multiple et analyse de la variance multiple – manova -).

La recodification

Les données disponibles ne se trouvent pas toujours dans un format adéquat. La recodification permet de générer de nouvelles variables mieux appropriées à l'analyse. On peut distinguer :

- La recodification d'une seule variable : mettre en classes une variable numérique, agréger les modalités d'une variable nominale, recoder une variable texte en fonction de son contenu.
- La recodification de plusieurs variables : calculer un score à partir des données de plusieurs variables nominales, calculer une nouvelle variable en application d'une formule en faisant intervenir plusieurs variables, créer de nouvelles variables calculées en application d'une analyse multivariée (scores factoriels, classifications...).

Les niveaux d'analyses

Analyser la population comme un tout: synthétiser

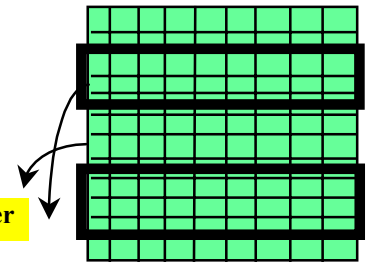
*On s'intéresse aux **variables***

Analyser les données individuelles: détailler

*On s'intéresse aux **individus***

Analyser les sous ensembles de la population: segmenter

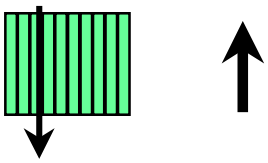
*On s'intéresse aux **strates***



Les problématiques

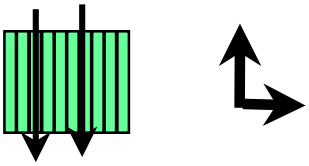
Analyse univariée

***Décrire** une variable à la fois*



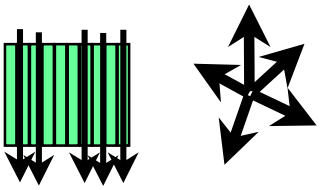
Analyse bivariée

*Mettre en relations 2 variables, **expliquer***



Analyse multivariée

Analyser simultanément
*plusieurs variables, **synthétiser***



L'analyse univariée

L'analyse univariée consiste à donner une description synthétique de l'ensemble des individus observés ou d'un de ses sous-ensembles. La manière d'effectuer ces descriptions dépend de la nature statistique des variables en jeu. Dans le cas d'étude par sondage, on peut également se livrer à l'estimation des propriétés de la population mère.

Décrire

On analyse chaque variable pour parvenir à un énoncé synthétique du type :

- ***" Il y a ....."*** : Il y a tant d'hommes, tant de femmes qui connaissent notre produit, il y a chez les acheteurs tant d'ouvriers...
- ***" ..... est (sont).... "*** : Le revenu moyen des ménages acheteurs est de.., la qualité est le premier critère de choix...
- ***".....fait (font)....."*** X% de l'échantillon fait confiance à.....

Ces énoncés spécifient l'état d'une variable par un effectif (tant d'hommes), un pourcentage (X% de la population fait) une grandeur moyenne (le revenu moyen des ménages acheteurs est...) ou un rang (le premier critère de choix est la qualité). L'importance du phénomène considéré se trouve ainsi chiffrée. L'analyste peut, sur cette base, exercer son jugement pour décider. Si la précision le permet, il pourra effectuer des calculs et des prévisions.

Dans le cas d'une variable texte, la description consiste à restituer le texte sous forme de liste ou lexique indiquant la fréquence des termes utilisés. Dans tous les cas, il faudra lire et interpréter ces éléments dont le sens n'est pas, comme pour les autres variables, fixé a priori.

Les indicateurs selon le type de variable

Selon le type de variable, la nature de l'information recueillie n'est pas de même nature et n'autorise pas les mêmes traitements. A chaque cas correspondent des indicateurs de résultats différents.

Dans le tableau ci-contre, on passe d'une information très bien structurée (variables numériques) à une information floue et ambiguë (variables textes), les opérations auxquelles elles se prêtent vont du calcul algébrique (variables numériques) à la quête du sens (variables textes). En conséquence, la qualité des indicateurs résultant et le "rendement" des méthodes mises en œuvre va décroissant. Ces propriétés déterminent des techniques de calcul ou de traitement de l'information, elles devraient également guider, dès la conception du questionnaire, le choix des questions.

Extrapoler à la population totale

Si l'échantillon répond aux conditions du sondage aléatoire (tous les individus de la population ont exactement la même chance d'être sélectionnés) et s'il comporte au moins 30 individus, on peut estimer à partir des indicateurs calculés sur l'échantillon, les propriétés de la population totale. La valeur de l'écart-type (numérique) et celle de l'intervalle de confiance (nominale) donnent les fourchettes de l'estimation.

Attention, dans le cas d'une strate, c'est l'effectif de la strate qu'il faut considérer et non pas celui de l'échantillon.

Le type de variable détermine les possibilités d'analyse

Type de variable	Nature de l'information	Opérations possibles	Indicateurs
numérique	nombre	calcul : +, -, x, : >, <, =, #	moyennes, écarts-types, effectifs, pourcentages
échelle	rang, ordre	comparaison : >, <, =, #	effectifs, pourcentages, sous certaines hypothèses moyennes et écarts-types
nominale	code : suite de caractères affectés d'une signification	identification : =, #	effectifs, pourcentages
texte	forme graphique : suite de caractères sans signification a priori	interprétation : ?	recodification, effectifs ou pourcentages de forme graphiques

L'analyse bivariée

L'analyse bivariée commence par la formulation d'une hypothèse orientée par la signification des variables et se poursuit par la mise en œuvre d'une méthode résultant de la nature des variables.

La sémantique : explication et relations causales

L'analyse bivariée conduit à formuler un énoncé de type Si..V_1.. Alors $V_{2...}$ postulant l'hypothèse d'une relation causale entre 2 variables. A ce stade, c'est la sémantique qui guide l'analyse : la signification des variables conduit à formuler une théorie justifiant la relation et son sens. La statistique peut confirmer ou invalider son existence mais seule la théorie en donne le sens. Ainsi, on peut établir statistiquement un lien entre le niveau d'éducation et le revenu. Mais c'est en fonction d'une théorie qu'on interprétera cette relation pour dire que le revenu conditionne l'éducation (théorie du coût) ou que l'éducation détermine le revenu (théorie du rendement).

L'anlayse bivariée commence donc par la formulation des hypothèses que la statistique permettra de tester. Le modèle ci-contre peut orienter la réflexion : parmi toutes les relations envisageables, 3 renvoient à des théories très générales du comportement humain :

- 1 - Le déterminisme sociologique : l'action obéit aux habitudes et aux contraintes.
- 2 - La décision et la rationalité : l'action résulte des choix et des calculs.
- 3 - Les psychologismes : l'action est modalité d'expression.

On peut bien sûr faire l'économie de la réflexion préalable et essayer toutes les relations envisageables. Elles peuvent être très nombreuses et perdre l'analyste dans une quête aveugle. D'autre part, le fait de constater une relation statistique ne suffit pas à établir une connaissance argumentante. Les exemples sont nombreux dans les études, de coïncidences inexpliquées ou fortuites...

Avant de commencer toute analyse bivariée, il convient donc d'établir une stratégie de recherche, en mobilisant les expériences, intuitions, croyances, théories, toutes connaissances préalables que l'on confrontera aux informations contenues dans la base de données. On déterminera ainsi quelles variables mettre en relation.

La statistique conduira à rejeter l'hypothèse si on ne peut pas montrer que la relation recherchée existe. Dans le cas inverse, la qualité de l'interprétation du fait statistique ne dépendra que de celle de la théorie utilisée.

La statistique : mettre en œuvre la méthode adaptée

La méthode à mettre en œuvre pour tester l'existence d'une relation dépend de la nature des variables en présence.

Le tableau ci-contre identifie les 3 cas possibles selon la nature des variables (nominales ou numériques).

Pour les variables échelles, on se ramène à l'un des cas précédents selon qu'on les considère comme des variables numériques ou comme des variables nominales.

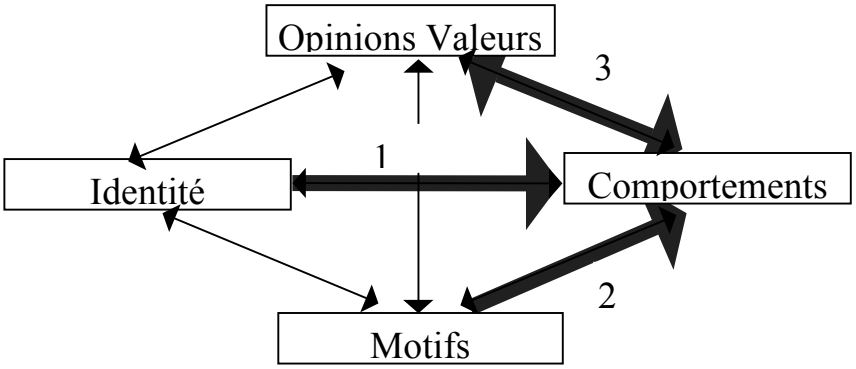
Le cas des variables textuelles peut également être ramené à l'un des 3 cas précédents :

- En faisant une analyse de contenu par recodification de la variable, on est ramené au cas du croisement avec une variable nominale.
- En utilisant les méthodes de l'analyse lexicale :

construction d'un tableau lexical qui décompte le nombre d'occurrences des mots de la variable texte analysée, pour les modalités d'une variable nominale. On est ramené au traitement d'un tableau de contingence analogue à ceux des tris croisés,

création de nouvelles variables d'origine lexicale : nominales (fermées sur les mots du lexique) ou numériques (longueur, banalité, intensité...) susceptibles d'être mises en relation entre elles ou avec les variables de contexte) ; ce qui ramène à l'un des cas précédents.

La sémantique : la signification des variables guide la formulation des hypothèses



La statistique : la nature des variables détermine la méthode

	V2 : NOMINALE	V2 : NUMERIQUE
V1 : NOMINALE	**1 Tris croisés** Recherche de correspondances par le calcul des effectifs relatif aux croisements des modalités de chaque variable (tableaux de contingence). Test du chi2 : analyse des correspondances. Séquences : ***Analyser / Tableaux croisés*** ***Analyser / Tableaux multiples***	**3 Analyse de la variance** Recherche des variations de V2 selon les modalités de V1 par le calcul des moyennes et écart-types de V2 sur les strates définies par les modalités de V1. Test de comparaison de moyennes, analyse de la variance. Séquence : ***Approfondir / Tableau de moyennes croisées***
V1 : NUMERIQUE	**3 Analyse de la variance** Recherche des variations de V1 selon les modalités de V2 par le calcul des moyennes et écart-types de V1 sur les strates définies par les modalités de V2. Test de comparaison de moyennes, analyse de la variance. Séquence : ***Approfondir / Tableau de moyennes croisées***	**2 Corrélation** Recherche d'une relation graphique ou algébrique entre les valeurs des 2 variables (droite et équation de régression). Test de corrélation : analyse de la relation. Séquence : ***Analyser / Corrélation et nuage de points***

L’analyse bivariée : Chi2 et AFC

Afin de déterminer s’il existe un lien entre 2 variables nominales, on compare les effectifs du tableau à ceux qu'on aurait obtenus si les effectifs étaient répartis proportionnellement. Si tel est le cas, quelle que soit la modalité considérée d’une des variables, la répartition des modalités de l’autre reste la même. Il y a alors indépendance entre les 2 variables.

Le calcul du Chi^2

Le test du Chi² consiste à déterminer si le tableau étudié correspond à cette hypothèse. S’il s’en écarte beaucoup, on présume l’existence d’un lien d’autant plus significatif que l’écart est grand. Le Chi² est la mesure de cet écart.

A partir d'un tableau de contingence à L lignes Li et C colonnes Cj, on détermine d’abord, pour chaque case, l’effectif théorique. C'est le produit du total de sa ligne (Li,.) par le total de sa colonne (C.,j) divisé par le total général (n) soit (Li,.*C.,j/n). On calcule ensuite la somme des carrés des écarts entre effectif réel et effectif théorique. Plus l’écart est grand, plus le Chi² est élevé, et plus on a de chances d’être en présence d’un lien significatif.

Cette appréciation dépend bien sûr de la dimension du tableau, c’est-à-dire du nombre de degrés de liberté : ddl = (L-1)* (C-1). Plus il y a de cases, plus la somme risque d’être élevée. Si la valeur du Chi² permet d’indiquer l’existence d’un lien, il faut, pour le qualifier, examiner comment celui-ci est composé.

Contribution au Chi² et interprétation des correspondances

Sur quelles cases observe-t-on les écarts les plus importants ? Sur quelles autres les effectifs sont sans surprise ?

On examine pour cela les contributions de chaque case à la somme du Chi², elles mettent en évidence l'importance de l'excès ou du déficit observable dans chaque cellule. Les cases contribuant le plus fortement sont encadrées de bleu ou de rouge selon que l'effectif réel excède ou est inférieur à l'effectif théorique.

Ainsi, c’est l’examen des contributions au Chi² et des correspondances qu’elles révèlent qui permet véritablement de qualifier la relation.

Construction et lecture d’une carte d’AFC

On peut donner une représentation plus visuelle des écarts à l’indépendance par la technique de l’analyse factorielle des correspondances. Elle conduit à tracer une carte qui dispose les modalités des 2 variables en fonction des écarts à la situation d’indépendance.

Par défaut, chaque modalité est représentée par un pavé de surface proportionnelle à son effectif. Leurs positions les unes par rapport aux autres s’interprètent ainsi :

- 2 modalités lignes et colonnes seront d’autant plus proches que les effectifs du tableau sont en excès par rapport à l’indépendance : attraction.
- Les modalités lignes et colonnes seront d’autant plus éloignées que les effectifs du tableau sont en déficit par rapport à l’indépendance : répulsion.
- Les modalités lignes ou colonnes situées à la périphérie de la carte signalent des profils originaux. Au contraire, une position centrale interdit tout commentaire (profils sans originalité ou point mal représenté dans le système d’axes de la carte).

Le bien fondé de ces interprétations dépend de :

- L’intensité du lien entre les 2 variables, mesuré par le Chi².
- La quantité d’informations restituée par la carte, indiquée par le pourcentage de variance expliquée (ou d’écart à l’indépendance) par les axes. La qualité de la représentation est d’autant meilleure que ces pourcentages sont élevés.
- L’interprétation des axes à partir des oppositions qu’ils mettent en évidence doit tenir compte du pourcentage de variance restituée. S’il est faible, il faut se garder d’insister sur des phénomènes qui ne représentent qu’une petite partie des caractéristiques du tableau.

Hypothèse : Le point de vue privilégié à l'achat a une influence sur la marque choisie. (Référence théorique : la décision rationnelle : Si Motifs ---> Comportement)

MARQUE x CRITERES

- Quelle est la marque de la voiture que vous utilisez ?
- Quels sont pour vous dans la liste suivante les trois principaux points à prendre en considération pour l'achat d'une automobile ?

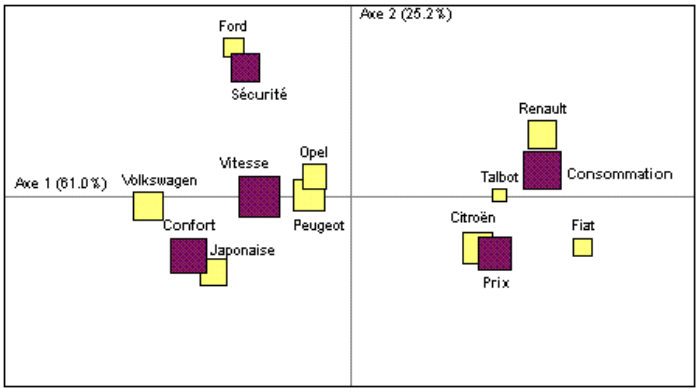
Tableau de contingence

CRITERES MARQUE	Vitesse	Confort	Sécurité	Consommation	Prix	TOTAL
Renault	22%	0%	17%	100%	17%	**100%**
Peugeot	64%	64%	36%	50%	29%	**100%**
Citroën	44%	17%	0%	56%	72%	**100%**
Talbot	25%	0%	25%	50%	75%	**100%**
Ford	50%	10%	80%	10%	10%	**100%**
Fiat	0%	17%	0%	83%	83%	**100%**
Volkswager	59%	55%	23%	0%	0%	**100%**
Opel	64%	36%	36%	36%	27%	**100%**
Japonaise	44%	61%	6%	6%	17%	**100%**
TOTAL	**45%**	**27%**	**22%**	**49%**	**35%**	**100%**

Test du chi2
Les données ne contredisent pas l'hypothèse

La dépendance est très significative. chi2 = 121,09, ddl = 32, 1-p = >99,99%.
Les cases encadrées en bleu (rose) sont celles pour lesquelles l'effectif théorique est nettement supérieur (inférieur) à l'effectif réel.
Attention, 26 cases ont un effectif théorique inférieur à 5, les règles du chi2 ne sont pas réellement applicables..

Analyse factorielle des correspondances



L'analyse bivariée : corrélation

Lorsque les variables sont numériques, on dispose d'une information très riche autorisant la recherche d'une formule mathématique pour qualifier la relation.

Nuage de points et droite de régression

La recherche d'une relation entre 2 variables numériques ***x*** et ***y*** peut se faire de 2 manières différentes :

- D'une manière graphique, en représentant chaque observation par ses coordonnées **x** et **y** selon 2 axes. On obtient alors un nuage de points plus ou moins bien alignés.
- D'une manière algébrique, en recherchant l'existence d'une relation linéaire entre ces 2 variables **y** = a**x** + b. On obtient des valeurs calculées **y** plus ou moins proches des valeurs observées.

Selon que la forme du nuage est plus ou moins proche d'une droite (la droite de régression), ou que les valeurs calculées à partir de l'équation (de régression) sont plus ou moins proches des observations réelles, on dira que la corrélation entre les 2 variables est bonne ou mauvaise. Le coefficient de corrélation mesure la qualité de l'ajustement entre les valeurs **y** et **x** réelles et le modèle de la relation représenté par l'équation **y** = a**x** + b ou par la droite correspondante. En référence à une interprétation causale du modèle, **y** est appelée variable à expliquer et **x** variable explicative.

Coefficient de corrélation

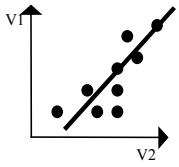
Le coefficient de corrélation (*r*) établit la qualité de l'ajustement entre deux variables V1 et V2. Sa valeur absolue (ou r^2) est comprise entre 0 et 1.

Elle est égale à 1 si l'ajustement est parfait : il existe une fonction V1=*a*xV2+*b* dont le résultat donne toujours exactement la valeur observée de V1 : on peut alors dire que V1 dépend exactement de V2.

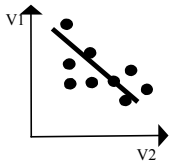
Elle est égale à 0 si quelle que soit la fonction considérée, la valeur calculée de V1 est également éloignée de sa valeur observée. V1 est indépendante de V2.

L'usage est de considérer qu'à partir d'un coefficient de corrélation de valeur absolue supérieure à 0.8, il existe une bonne relation entre les 2 variables.

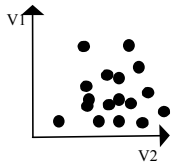
Le coefficient de corrélation est également affecté d'un signe. Il indique le sens de la relation. Elle est croissante si le signe est positif, décroissante sinon. Ce signe est aussi celui de *a*, le coefficient de régression dans l'équation V1=axV2+b.



Relation croissante, a>0, r^2>0.8



Relation décroissante, a<0, r^2>0.8

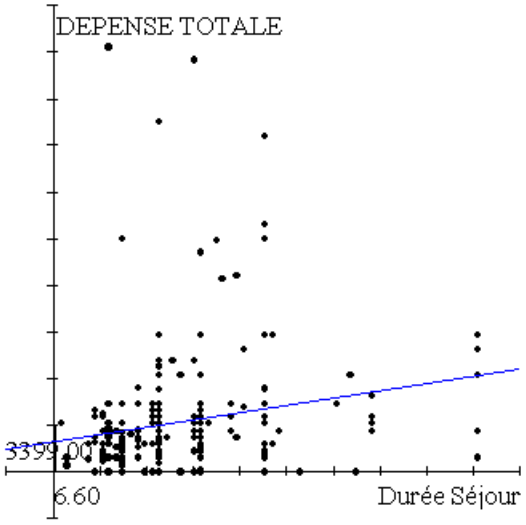


Pas de relation r^2<0.1

Dans le cas d'une régression multiple, on cherche à établir une relation du type V1=*a*xV2+*b*xV3+*c*xV4. On calcule alors un coefficient de régression multiple. Il indique également la qualité de l'ajustement effectué par le modèle et s'interprète comme un coefficient de corrélation simple.

Modèle linéaire et non linéaire

Le calcul du coefficient de corrélation se fait par rapport au modèle linéaire de l'équation du premier degré à laquelle correspond la droite de régression. On peut tester la relation par référence à un modèle non linéaire. Il faudra pour cela transformer la valeur de la variable expliquée en la recalculant par rapport au modèle à tester. Par exemple, pour tester un modèle logarithmique, on calculera une nouvelle variable égale au log de la variable explicative : log(x) pour ensuite tester l'equation y=a.logx+b.

Hypothèse : Plus le séjour est long, plus la dépense totale est élevée (Théorie comptable élémentaire de la proportionnalité des coûts).	Corrélation. DEPENSE TOTALE /. DUREE SEJOUR
Nuage de points et droite de régression	 *209 individus ne sont pas pris en compte (non-réponse à au moins un des critères).* *Chaque individu est représenté par un point.*
Equation de la droite de régression	*Le graphique montre les 349 points de coordonnées DUREE SEJOUR; DEPENSE TOTALE* Equation de la droite de régression : DEPENSE TOTALE = 81.05 * Durée Séjour + 2210.08
Test de corrélation Les données contredisent l'hypothèse	*Coefficient de corrélation : 0.198 (Durée Séjour explique 3 % de la variance de DEPENSE TOTALE)* *Ecart-type du coefficient : 21.588*

L'analyse bivariée : analyse de la variance

L'analyse de la variance s'applique au cas où les variables sont de natures différentes : l'une est nominale et définit des catégories d'individus, l'autre, numérique, permet de mesurer une propriété de ces individus. On pourra ainsi par exemple savoir si, selon le sexe, le revenu varie significativement.

Tableau de valeurs moyennes

La construction d'un tableau de valeurs moyennes, établit, pour les modalités d'une variable nominale, les valeurs moyennes d'une ou plusieurs variables numériques. On peut ainsi comparer ces valeurs entre elles et conclure à l'existence d'une relation si les variations de la moyenne mettent en évidence des différences significatives sur l'ensemble des modalités.

On utilise pour cela le test de l'analyse de la variance.

Analyse de la variance

Son but est d'établir si, au regard des valeurs de la variable numérique, les groupes d'individus correspondant aux modalités de la variable nominale sont significativement différents les uns des autres. Elle met en œuvre les principes suivants :

- Pour chaque modalité de la variable nominale, la moyenne de la variable numérique dissimule une dispersion autour de cette moyenne. La variance (le carré de l'écart-type) mesure cette dispersion appelée **variance interne**.
- D'une modalité à l'autre, la moyenne varie et révèle une hétérogénéité, plus ou moins grande, mesurée par un autre calcul de variance : **la variance entre modalités.**

Le test d'analyse de variance porte sur le rapport entre la variance entre modalités et la moyenne des variances internes. Il a pour but de vérifier si l'hétérogénéité entre modalités est plus grande que l'hétérogénéité à l'intérieur des modalités. Si ce rapport noté F est suffisamment élevé, on dit que la variable numérique discrimine les modalités de la variable nominale.

Ce jugement global s'applique à la répartition de la population totale en catégories définies par chacune des modalités. On peut le compléter en comparant les valeurs moyennes de chaque modalité à la moyenne établie sur l'ensemble de la population.

Critères discriminants et valeurs significatives

Les tests d'analyse de la variance et de comparaison de moyennes apportent une information très utile permettant d'identifier les variables pour lesquelles les catégories de la variable nominale font apparaître des différences discriminantes :

- Les variables dont le nom est encadré de bleu discriminent les modalités de la variable nominale. Pour ces variables, le test de Fisher est significatif (par défaut au risque de 5 %).
- Les cellules encadrées dans le tableau signalent une moyenne significativement différente de la moyenne sur l'ensemble de l'échantillon. Le test de comparaison de moyennes est significatif (par défaut au risque de 5 %).

Représentation graphique de la dispersion

L'analyse d'un tableau de moyennes est facilitée par la représentation graphique de la dispersion des variables numériques selon les modalités de la nominale.

Analyse de la variance à plusieurs facteurs

On peut mener une analyse de la variance en considérant 2 variables nominales et une numérique. On cherche alors, par la comparaison de tous les cas définis par les nominales les influences directes et croisées qu'elles peuvent avoir sur la valeur de la numérique. Cette méthode dite analyse de la variance multiple (Manova) fait partie des méthodes multivariées. Elle n'est disponible que dans Plus^2.

Hypothèse : Le mode d'hébergement a une influence sur les dépenses et la durée du séjour (référence au mode de production et élasticité prix).

Tableau de moyennes de.MODE HEBERGEMENT

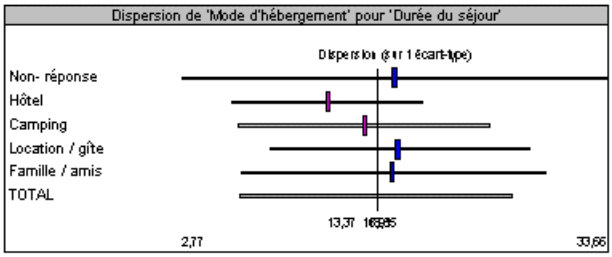
Critères évalués : DUREE DU SEJOUR, DEPENSE TOTALE

Tableau des valeurs moyennes pour Durée du séjour Dépense totale et Dépense d'habillement.

Mode d'hébergement	Durée du séjour	Dépense totale
Hôtel	13,37	5007,26
Camping	16,00	2536,15
Location / gîte	18,55	4643,26
Famille / amis	18,16	2803,65
TOTAL	16,91	3588,18

Les valeurs du tableau sont les moyennes calculées sans tenir compte des non-réponses.
Les noms des critères discriminants sont encadrés.
Les nombres encadrés correspondent à des moyennes par catégorie significativement différentes (test t) de l'ensemble de l'échantillon (au risque de 5%).

Graphique de dispersion



Test d'analyse de la variance : l'hypothèse est confirmée : le mode d'hébergement a une incidence sur la durée du séjour et sur la dépense.

Résultats du test de Fisher :
Durée du séjour : V_inter = 505,59, V_intra = 94,96, F = 5,32, 1-p = 99,96%
Dépense totale : V_inter = 101930198,55, V_intra = 14443866,90, F = 7,06, 1-p = >99,99

Analyse multivariée : objectifs et méthodes

L'analyse multidimensionnelle des données consiste à analyser simultanément plus de deux variables à la fois dans un but de synthèse ou d'analyse.

Synthétiser

Il s'agit de résumer la masse des informations concernant un grand nombre d'individus décrits par de nombreuses variables.

On cherche à décrire les données par une expression plus économique afin d'en rendre compte plus simplement. Deux voies sont envisageables :

- **La classification ou la typologie :** elle consiste à regrouper les individus de la base de données en classes (ou types) d'individus selon les similarités qui permettent de les rassembler. Par exemple : décrire une population en identifiant différentes classes de styles de vie.
- **La réduction des dimensions d'analyse :** elle consiste à réduire un grand nombre de variables à quelques dimensions communes. Par exemple, ramener la variété des opinions exprimées par les réponses à vingt questions différentes, à 2 dimensions opposant d'une part, le sentiment à la raison, et d'autre part, la tradition au changement.

Analyser

On cherche les influences de plusieurs variables entre elles pour mettre en évidence comment celles-ci se déterminent les unes et les autres.

On cherche à expliquer en construisant des modèles permettant d'expliquer une variable par les variations de plusieurs autres, en cherchant ainsi des liens de causalité plus complexes que la simple relation entre deux variables.

Les méthodes

Les méthodes disponibles sont nombreuses :

Certaines sont directes. Elles peuvent consister à :

- combiner entre elles plusieurs variables pour calculer une nouvelle variable qui les résume (une somme, une moyenne, un score...).
- marquer des individus selon leur appartenance à telle strate ou selon qu'ils répondent à tel profil et constituer ainsi des groupes a priori.

D'autres sont indirectes. Elles reposent sur une analyse préalable de la structure des données orientant la synthèse en fonction des propriétés révélées. On peut distinguer ces méthodes selon le type de variables auxquelles elles s'appliquent :

- Pour les variables quantitatives, il s'agit de :

- **La corrélation multiple** : elle permet de construire la matrice de corrélation qui met en évidence l'ensemble des relations existant entre plusieurs variables.
- **L'analyse en composantes principales** : à partir des propriétés de la matrice de corrélation, elle permet de substituer aux variables de départ un plus petit nombre de dimensions qui les résument au mieux. Dans cet espace réduit, on peut alors définir une typologie en regroupant les individus en classes homogènes.
- **La régression multiple** : à partir des propriétés de la matrice de corrélation, elle permet de bâtir un modèle explicatif donnant la meilleure expression de l'une des variables - variable à expliquer - en fonction des autres - variables explicatives -.
- **La classification automatique** : cette méthode consiste à répartir les individus d'une population en un nombre de classes déterminé a priori. En fonction de la structure des données, la méthode consiste à améliorer une partition initiale des individus.

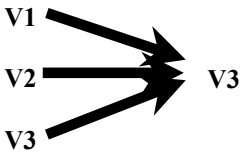
- Pour les variables qualitatives, il s'agit de :

- **L'analyse factorielle des correspondances multiples** : à partir d'une analyse des distances entre individus, définie par rapport à leur description sur un grand nombre de variables qualitatives, on détermine un sous-ensemble de dimensions, des facteurs conservant au mieux les distances de départ. Dans cet espace réduit, on peut alors définir une typologie en regroupant les individus en classes homogènes.
- **La classification automatique** appliquée à des variables qualitatives.

Toutes ces méthodes sont accessibles par le menu ***Approfondir***. On peut pour mémoire y rajouter des méthodes de même nature, ne portant pas sur les individus mais sur des tableaux d'effectifs ou des tableaux de moyennes. Le tableau ci-dessous résume les méthodes disponibles.

Expliquer

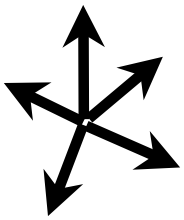
Etablir un modèle des relations entre variables. Il exprime ces relations sous forme de fonction mathématique.

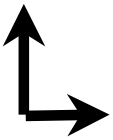


Régression multiple : n variables numériques
Manova : 2 variables nominales et 1 numérique

Synthétiser

Réduire le nombre des dimensions d'analyse : cartes factorielles, scores factoriels.

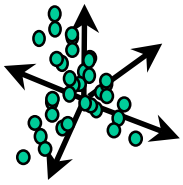


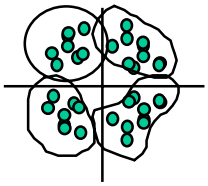


Analyse factorielle multiple (**AFCM**) : n variables nominales
Analyse en composantes principales (**ACP**) : n variables numériques
Analyse factorielle des correspondances (**AFC**) : 2 nominales
Analyse en composantes principales
d'un tableau de moyennes (**ACP**) : 1 nominale n numériques

Classifier

Regrouper les individus en classes homogènes. Classification automatique, typologie à vue.





Classification automatique : n numériques
Typologie AFCM : n nominales
Typologie ACP : n numériques

L'analyse factorielle : les principes

Les méthodes d'analyse factorielle peuvent porter sur des données de dimensions modestes, tableaux croisés, tableaux de moyennes… ou sur des ensembles de grande dimension lorsqu'en ligne, on trouve tous les individus de la population. Dans ce dernier cas, la recherche des facteurs peut se prolonger par la construction d'une typologie.

AFCM et ACP

Les deux méthodes dont on présente ici les principes portent sur ce type de données comportant autant de lignes que l'échantillon ou la strate compte d'individus.

Quant aux colonnes, ce sont :

- soit des variables numériques ou critères (questions ouvertes numériques ou échelles), auquel cas, il s'agit d'un tableau de valeurs (chaque case est la valeur donnée par l'individu en réponse à la question posée) et l'analyse est une **Analyse en Composantes principales** ou ACP sur individus.
- soit des variables qualitatives (questions fermées uniques ou multiples), auquel cas, elles représentent les modalités. Le tableau est alors un tableau binaire ne comportant que des 1 ou des 0, selon que l'individu a cité ou non la modalité en question. Un tel tableau binaire peut être assimilé à un tableau d'effectifs et peut être soumis comme tel à l'**Analyse factorielle des correspondances multiples** ou AFC sur individus.

Bien qu'elles concernent des variables de natures différentes, ces deux méthodes mettent en œuvre les mêmes principes et la même démarche. Nous les décrivons d'une manière commune dans ce qui suit, pour illustrer plus en détail ensuite l'***Analyse en composantes principales*** et situer enfin, par différence, les spécificités de l'***Analyse factorielle multiple***.

Les données individus / variables

Le tableau de *i* lignes, représentant les individus, et de *c* colonnes, représentant les dimensions, correspond à l'univers à analyser. Il forme un hyper-espace de *c* dimensions dans lequel se situent les *i* individus.

Les techniques qui nous intéressent ont pour but de ramener cet espace à de plus modestes dimensions.

L'idéal est de le réduire aux deux dimensions d'un plan. Cette configuration permet en effet de visualiser les individus et de mettre en évidence ce qui les distingue. Ces deux dimensions, appelées facteurs, résument l'ensemble des variables de départ et peuvent conduire à une interprétation plus synthétique de la structure des données.

A partir de ces mêmes informations, on peut regrouper les individus en classes de proximité pour définir ainsi une typologie.

Cette démarche est illustrée par le schéma ci-contre.

Trouver les facteurs pour réduire le nombre des variables d'analyse

La réduction de l'espace de départ se fait par la recherche des facteurs résumant l'ensemble des données.

Pour l'***Analyse en composantes principales*** - cas des données quantitatives -, on cherche les combinaisons linéaires des variables qui conservent le mieux la structure des données. On s'appuie pour cela sur l'analyse de la matrice de corrélation.

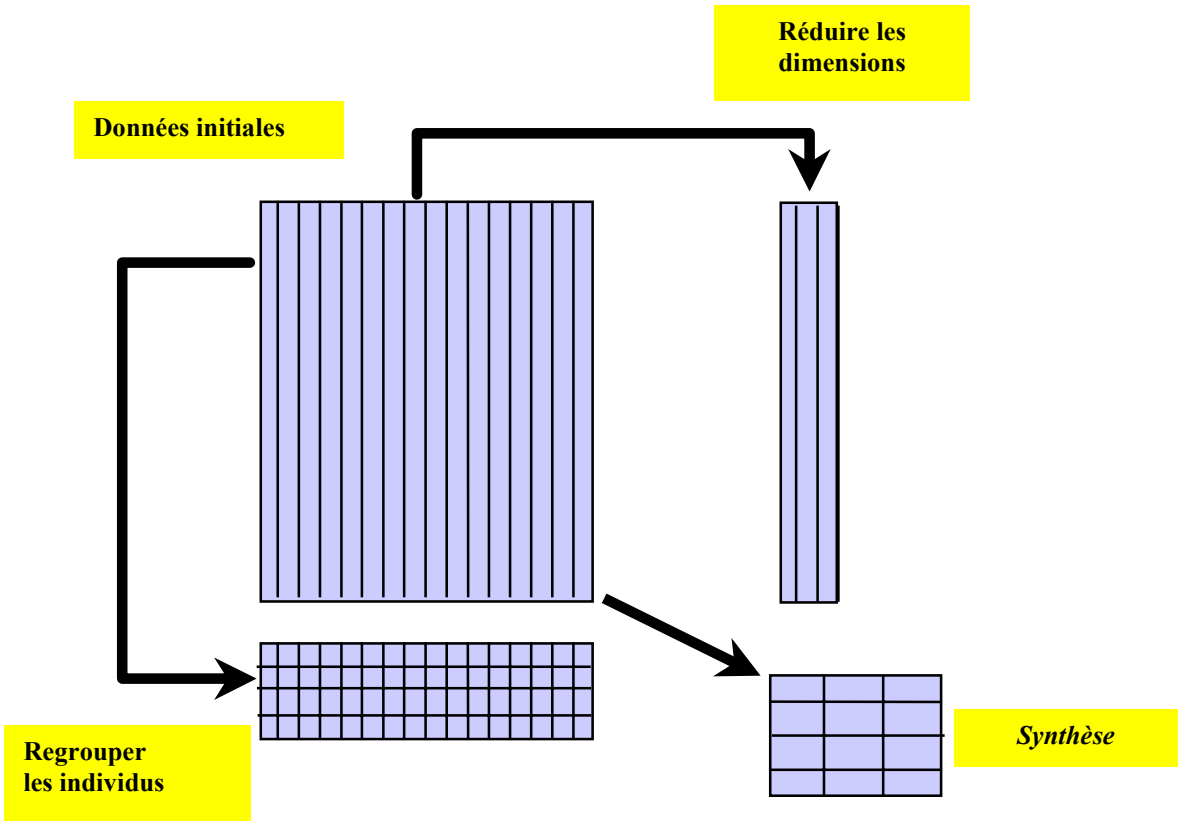
Pour l'***Analyse factorielle des correspondances*** – cas des données qualitatives -, on procède par une analyse de la distance du Chi² en vue de déterminer les facteurs qui restituent au mieux les écarts entre individus. On s'appuie pour cela sur le tableau des effectifs ou tableau de Burt.

Agréger les individus : construire une typologie

La détermination des facteurs résumant l'information de départ permet, par projection, de représenter les individus dans un plan. Les individus apparaissent ainsi sous la forme d'un nuage de points mettant en évidence différentes zones. On peut regrouper les individus en catégories selon leur disposition sur la carte et définir ainsi une typologie.

Scores factoriels et typologie

L'enregistrement des coordonnées des individus par rapport aux axes factoriels (score factoriels) et leur appartenance à l'une des catégories de la typologie permet de résumer l'information de départ.



Variables numériques : **Analyse en composantes principales**

Variables nominales : **Analyse factorielle multiple**

On enrichit la base de départ en créant
de nouvelles variables correspondant aux :

Scores factoriels
Classes typologiques

L'analyse factorielle : interpréter les résultats

Qualité de la réduction par l'analyse factorielle

Les algorithmes utilisés consistent à rechercher les n premiers facteurs. Chacun d'eux restitue une partie de l'information du tableau de départ. Le premier facteur en restitue le plus et ainsi de suite, selon l'ordre décroissant de la quantité d'informations.

La qualité de l'analyse dépend ainsi de l'information restituée par les deux premiers facteurs. Plus celle-ci est élevée, meilleure sera l'analyse effectuée dans le plan qu'ils définissent. Lorsque celle-ci est trop faible, il peut être nécessaire de la compléter par l'analyse des plans suivants.

La quantité d'informations restituée par un facteur (ou axe factoriel) est indiquée par le pourcentage de variance expliquée par le facteur (AFC) ou la composante (ACP).

Interpréter les facteurs

L'interprétation des facteurs peut se faire par l'examen du plan factoriel ou à partir du tableau des contributions. Les contributions indiquent dans quelle mesure les axes restituent l'information contenue dans les variables de départ.

Visuellement, on peut raisonner comme si, par leur position éloignée du centre, les variables (ACP) ou les modalités (AFCM) « tiraient les axes factoriels en leur donnant leurs propriétés ». Au contraire, lorsqu'elles se trouvent près du centre, elles n'ont pas ou peu d'influence.

Dans le cas de l'ACP, les coordonnées des variables sont égales au coefficient de corrélation de la variable avec chacun des axes. Pour l'AFCM, elles sont un indicateur de la contribution des modalités à chacun des axes.

Trouver une typologie pour grouper les individus en classes homogènes

Par projection, on peut représenter les individus dans le plan factoriel, mais selon leur position dans l'espace d'origine, ils seront plus ou moins bien représentés. S'ils sont éloignés du plan de projection, ils se projetteront au centre du plan et leur position fera illusion.

Dans le cas de l'ACP, les coordonnées des variables sont égales au coefficient de corrélation de la variable avec chacun des axes. Pour l'AFCM, elles sont un indicateur de la contribution des modalités à chacun des axes.

Ainsi dans la figure suivante, les individus C et A semblent proches alors qu'ils sont en fait éloignés. A est mal représenté car très éloigné du plan factoriel avec lequel il forme un angle presque droit.

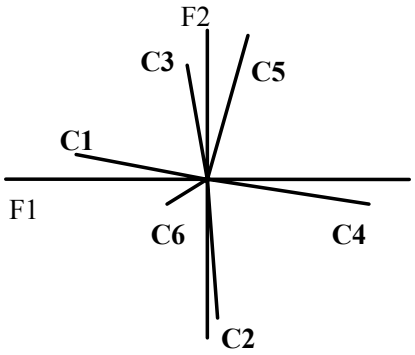
La projection est déformante, il convient donc de procéder avec prudence, par exemple en ignorant les points mal projetés (ceux qui font un angle élevé avec le plan) et en concentrant l'analyse sur les groupes de points éloignés du centre.

Ces précautions étant prises, il est alors possible de regrouper les individus, selon leur proximité dans le plan factoriel, en construisant ainsi une typologie à vue. Cette possibilité est une des originalités du Sphinx. Elle est accessible à partir du bouton ***Typologie*** figurant en regard des cartes d'analyses factorielles.

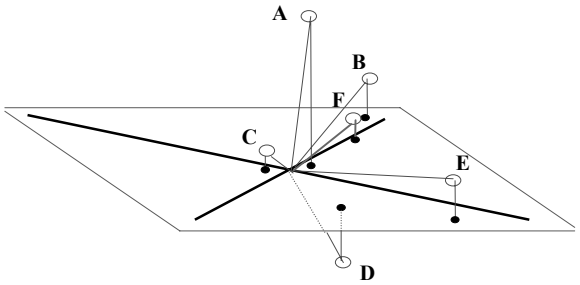
Interpréter les axes

Le schéma ci-contre conduit à interpréter le facteur F1 comme la dimension sur laquelle s'opposent les critères C1 et C4, alors que sur le facteur F2, C3 et C5 s'opposent à C2. Les variations du critère C6 sont mal représentées par ce plan factoriel.

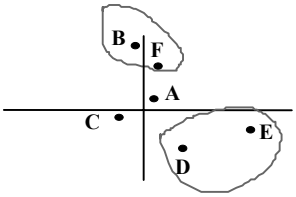
L'examen du tableau des contributions associé à cet exemple confirmerait ces interprétations. Par exemple, on pourrait y vérifier que C6 ne contribue que pour un très faible pourcentage à chacun des axes.



Projection des individus dans le plan factoriel



Vue dans l'espace



Vue dans le plan factoriel

Corrélation et régression multiple

L'objectif de la régression multiple est de mettre en relation une variable, la variable expliquée avec plusieurs autres les variables explicatives, dans le but de définir l'équation qui les relie. On pourra ainsi, connaissant les variables expliquées, déterminer les valeurs de la variable explicative : V1=*a*xV2+*b*xV3+*c*xV4. On calcule alors un coefficient de régression multiple. Il indique la qualité de l'ajustement effectué par le modèle et s'interprète comme un coefficient de corrélation simple.

L'exemple ci-contre, tiré d'une étude de satisfaction, montre comment on peut expliquer la satisfaction globale en la rapportant à l'évaluation des différents attributs du service considéré. Les coefficients de régression partielle (a_i) donnent une indication sur la manière dont la perception de chaque attribut intervient dans la formation de la satisfaction globale.

Equation de régression linéaire multiple et paramètres d'ajustement

On calcule l'équation linéaire qui ajuste le mieux la variable expliquée par rapport aux variables explicatives. Les résultats sont communiqués sous la forme de l'équation de régression multiple.

La qualité de l'ajustement s'apprécie principalement à la valeur du coefficient de corrélation. Plus sa valeur absolue est élevée, plus faible est l'écart entre les valeurs calculées et observées (cet écart est aussi appelé résidu).

L'effet de chaque variable explicative dépend des coefficients de régression figurant dans l'équation. Plus celui-ci est grand, plus la variable explicative considérée influence la variable expliquée. Mais il faut également tenir compte de l'écart-type de chacun de ces coefficients. Plus il est élevé, moins nette est l'influence de la variable considérée.

Enfin, l'indicateur F est un autre moyen d'apprécier la qualité de l'estimation. Si sa valeur est supérieure au seuil d'une table de Fisher, l'estimation est considérée comme très significative (à 95%) ou peu significative (entre 80 et 95%), sinon, elle ne l'est pas du tout.

Les variations de F peuvent conduire à reconsidérer les variables intervenant dans le modèle. En effet, la suppression ou l'ajout de variables supplémentaires peut affecter la qualité de l'ajustement (coefficient de corrélation) mais dégrader celle de l'estimation. On observe alors une décroissance de F.

Procédure de régression pas à pas

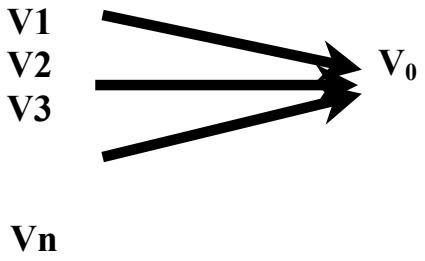
Les variables explicatives interviennent dans le calcul, dans l'ordre de leurs corrélations respectives avec la variable à expliquer. L'algorithme s'arrête quand la qualité de l'ajustement n'augmente plus de manière significative.

Indépendance des variables explicatives

L'examen de la matrice des coefficients de corrélation permet de vérifier que les variables explicatives intervenant dans l'équation sont indépendantes. On peut, en outre, tester la sensibilité de la régression en supprimant ou rajoutant des variables. On peut alors suivre l'évolution du coefficient de corrélation multiple.

Régression non linéaire

Il est tout à fait possible de tester des modèles non linéaires. Il faut pour cela au préalable transformer les variables intervenant dans le modèle en utilisant la fonction.



Variable expliquée

Variables explicatives

$$Vo = a_1xV_1 + a_2xV_2 + a_3xV_3 \ldots + a_nxV_n$$

Note *Evaluations*

Coefficients d'importance relative

Equation de régression multiple

note_globale = +0.970 * Restauration +0.891 * Confort -0.652 * Soins_qualité +0.639 * Informations -0.510 * Réconfort +0.390 * Hygiène -0.219 * Services +9.709

Matrice de corrélation

	note_glo	Restaura	Confort	Hygiène	Informati	Soins_ra	Compéte	Disponib	Réconfo	Accueil	Services	Soins_q
note_globale	1,00											
Restauration	0,42	1,00										
Confort	0,35	0,17	1,00									
Hygiène	0,26	0,33	0,09	1,00								
Informations	0,23	0,12	0,05	0,12	1,00							
Soins_rapidité	0,15	0,22	0,16	0,19	0,17	1,00						
Compétence	0,15	0,22	0,25	0,08	0,08	0,17	1,00					
Disponibilité	0,14	0,26	0,17	0,22	0,22	0,31	0,15	1,00				
Réconfort	-0,00	0,15	0,18	0,29	0,12	0,18	0,14	0,20	1,00			
Accueil	-0,06	-0,04	-0,02	-0,03	0,02	-0,09	-0,00	-0,06	0,03	1,00		
Services	-0,06	0,04	-0,03	0,18	-0,00	0,03	-0,02	0,04	0,00	0,02	1,00	
Soins_qualité	-0,14	0,15	0,10	-0,04	0,07	0,09	0,08	0,08	0,01	-0,04	0,04	1,00

Classification automatique

La classification automatique répond à l'objectif suivant : regrouper les individus d'une population en classes formant une partition. C'est-à-dire que tout individu appartient à une classe et à une seule. La partition est l'ensemble des classes.

Différentes approches de classification

Il existe plusieurs manières d'opérer une classification :

- effectuer le tri à plat ou croisé d'une ou plusieurs variables fermées uniques. Chaque classe est formée des individus ayant en commun une modalité (tri à plat) ou un couple de modalités (tri croisé). Toute variable fermée unique (ou toute combinaison de variables fermées uniques) définit ainsi une classification ;
- construire une typologie à vue à partir de l'examen d'un plan factoriel : on définit la variable fermée sur la typologie qui, comme toute variable fermée, définit une classification.

Dans le premier cas, la classification est la traduction exacte des données. Mais si on veut la construire à partir de plusieurs variables, on risque d'obtenir un nombre beaucoup trop grand de classes, ce qui fait perdre tout intérêt à la classification.

Dans le deuxième cas, on peut fixer un nombre réduit de classes ou types, pour capter les individus apparaissant dans un plan factoriel mais le procédé d'affectation peut paraître trop subjectif et approximatif.

La procédure de classification automatique proposée est un compromis pour éviter les inconvénients qui viennent d'être signalés. Elle repose sur la mise en oeuvre de la méthode dite des centres mobiles ou des nuées dynamiques.

Il existe bien d'autres méthodes (classifications hiérarchiques ascendantes ou descendantes) mais celle que nous proposons peut s'appliquer à de très grands effectifs et permet de contrôler a priori le nombre de classes.

Définir une classification

Les étapes de la méthode sont les suivantes :

- Fixer les objectifs de la classification ;
- Définir la population : tous les individus de la base ou une strate seulement ;
- Sélectionner les variables par rapport auxquelles on souhaite définir une partition de la population ;
- Fixer le nombre de classes désirées ou une partition de départ définie par la variable fermée unique de son choix ;

Il est tout à fait possible de choisir pour variable de départ une typologie définie à partir d'une analyse en composantes principales ou une analyse factorielle multiple. On pourra ainsi affiner les agrégations effectuées dans le plan factoriel.

La méthode des centres mobiles

Cet algorithme met en oeuvre le principe suivant : on améliore progressivement la partition de départ en calculant, pour chaque classe, un *individu fictif moyen*. On définit alors une nouvelle classification, par agrégation aux individus fictifs moyens résultant de la classification précédente. On répète l'opération tant que la nouvelle classification diffère de la précédente.

L'individu fictif moyen est calculé comme le barycentre de la classe, l'agrégation se fait en fonction de la moindre distance.

Analyser la classification obtenue

La classification résultant de l'algorithme dépend de la partition de départ. Elle est fixée d'une manière aléatoire si on se borne à indiquer le nombre de classes désirées, sinon, elle dépend de la variable choisie.

Il peut alors être intéressant de rechercher plusieurs classifications pour les comparer entre elles. On peut pour cela considérer :

- la répartition des effectifs conduisant à des classes plus ou moins équilibrées ;
- le pouvoir discriminant de la classification. On cherchera pour cela à comparer les classes du point de vue des variables à partir desquelles on les a déterminées : analyse de la variance ou test du Chi² selon qu'il s'agit de variables quantitatives ou qualitatives.