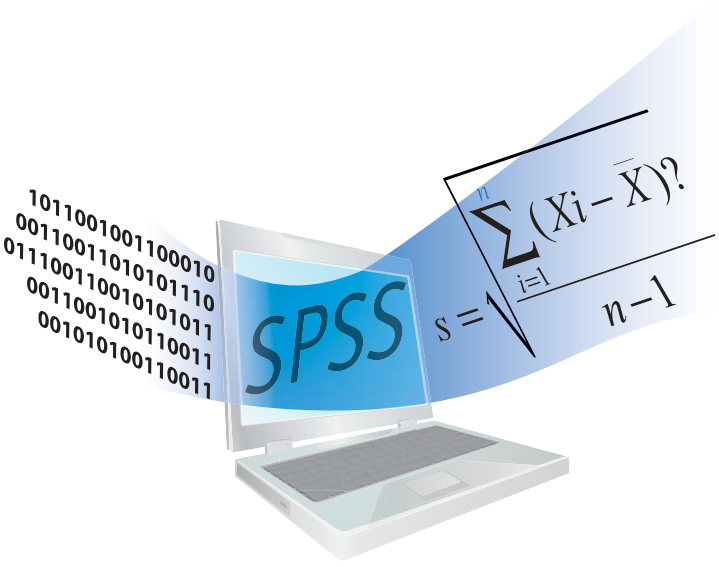
Sciences de gestion

Synthèse de cours & exercices corrigés

Analyse de données avec SPSS®



- Toutes les étapes clés d'une analyse de données
- Une pédagogie active avec le logiciel SPSS
- Les fichiers des exercices disponibles à l'adresse www.pearson.fr

collection Synthex

PEARSON Education

Manu CARRICANO
Fanny POUJOL

Sciences de gestion

Synthèse de cours & exercices corrigés

Analyse de données avec SPSS®

Manu Carricano
INSEEC Paris

Fanny Poujol
IAE Valenciennes

Directeur de collection : Roland Gillet
Université Paris I Panthéon-Sorbonne

Avec la contribution de Laurent Bertrandias pour la relecture de fond
Institution d'Administration des Entreprises – Université Toulouse 1

collection Synthex

PEARSON Education France

ISBN : 978-2-7440-4075-7
ISSN : 1768-7616

Mise en page : edito.biz

Sommaire

Préface

Il existe aujourd'hui de nombreux livres consacrés aux études de marché et à l'analyse marketing, ce que les Anglo-Saxons nomment *marketing research*. L'ouvrage de Fanny Poujol et Manu Carricano, *Analyse de données avec SPSS*, se distingue clairement de ceux existant sur le marché par son parti pris résolument opérationnel. L'instrumentation de gestion est souvent négligée dans la littérature francophone consacrée au management. Tendance bien cartésienne à la conceptualisation ? Bien des manuels, peut-être en contradiction avec la définition même de ce genre d'écrit, consacrent la portion congrue aux outils et à leur application pratique.

Analyse de données avec SPSS prend le contre-pied d'une telle tendance. Peut-être est-ce en raison de la jeunesse et de la singularité des auteurs : une docteur en gestion, aujourd'hui maître de conférences à l'université de Valenciennes, et un professeur assistant dans une grande école, tous deux chercheurs à l'Inseec et encore très proches des difficultés d'apprentissage de la recherche quantitative en marketing ?

En tout état de cause, les huit chapitres de leur ouvrage présentent avec rigueur les problèmes opérationnels de la recherche quantitative en marketing et leur résolution pratique, en prenant patiemment le lecteur par la main grâce à des exemples et des exercices et en le guidant dans l'utilisation du logiciel SPSS. Le titre des chapitres résume ce parcours initiatique dans la recherche quantitative en marketing : analyser pour décider, décrire les données, simplifier les données, segmenter, l'analyse de variance, la régression linéaire, l'analyse conjointe, communiquer les résultats. Dans ce parcours en huit étapes, c'est résolument l'application et la mise en œuvre pratique qui sont privilégiées aux dépens des considérations théoriques résumées clairement en tête de chapitre.

Outre l'aspect opérationnel, les auteurs ont également eu comme objectif de privilégier la dimension synthétique de leurs démonstrations. Il n'était pas question pour eux de faire une « somme » de plus sur le sujet, les bibliothèques étant déjà fournies en livres de ce type. Certes, d'autres méthodes mériteraient de figurer dans l'ouvrage, mais on ne peut reprocher aux auteurs d'avoir privilégié les techniques les plus couramment utilisées.

Nul doute que ce livre original connaîtra le succès qu'il mérite auprès des nombreux chercheurs en marketing, professionnels et universitaires. Il sera alors temps pour les auteurs d'offrir un second tome dans la même veine. C'est tout ce que nous leur souhaitons, pour eux-mêmes et pour leurs futurs lecteurs.

Pierre-Louis Dubois, Professeur
Université Panthéon-Assas (Paris II)
ESCP-EAP
Président du Comité scientifique de l'Inseec

Introduction

Approche adoptée

La plupart des décisions de l'entreprise reposent sur des données collectées sur le marché, les clients, les concurrents. Mais le plus souvent, ces analyses sont simplistes, limitées, voire biaisées : d'une part, parce qu'elles se limitent à des analyses descriptives (tableaux croisés, analyses factorielles…) et non pas explicatives des phénomènes observés ; d'autre part, parce qu'elles ne s'assurent pas toujours des conditions de validité et de fiabilité des résultats. Dans le même temps, l'exercice du marketing s'est considérablement transformé ces dernières années : le volume de données disponible est plus important, et les outils d'analyse plus sophistiqués. Ces solutions analytiques, telles les suites de logiciels développées par SPSS, visent à tirer parti de cette profusion de données afin d'aider les dirigeants à prendre des décisions fondées, optimales. Comme le signale Sunil Garga, président d'IRI Analytic Insight Group : « *Les approches analytiques en marketing ont amené à plus de changement durant les 24 derniers mois que lors de ces 24 dernières années.* »

La diffusion de ces nouvelles approches dans l'entreprise passe par la mise sur le marché de jeunes diplômés éclairés et sensibilisés à une démarche analytique dépassant l'intuition et fondée sur des modèles afin de prendre des décisions optimales. L'idée force qui nous a guidés tout au long de la rédaction de cet ouvrage est de démontrer la valeur ajoutée de l'analyse de données dans l'optimisation de décisions courantes au sein de l'entreprise. Le marketing, notre domaine de spécialisation, nous a semblé tout indiqué pour mettre en œuvre une telle approche fonctionnelle. Loin d'un inventaire de techniques statistiques, nous avons souhaité définir une série de questions simples faisant le lien entre les outils d'analyse de données et les décisions marketing, parmi lesquelles :

- Comment transposer un problème managérial en modèle d'analyse ?
- Comment améliorer la validité et la fiabilité d'un questionnaire ?
- Quelle approche mobiliser pour décrire les résultats d'une enquête ?
- Comment synthétiser les données collectées ?
- Comment segmenter un marché ?
- Comment expliquer l'influence d'une décision sur un marché ?
- Comment modéliser un comportement d'achat ?

- Comment expliquer les préférences des consommateurs ?
- Comment communiquer les résultats ?

Cet ouvrage s'adressant principalement à des étudiants – et s'inspirant fortement des remarques de nos propres étudiants –, nous avons également cherché à présenter les informations de manière simple, passant rapidement le relais à une mise en application des concepts statistiques par le biais d'une manipulation du logiciel SPSS. Apprendre en faisant permettra au lecteur d'acquérir des compétences en analyse de données de manière progressive, et sur l'ensemble de la démarche. Cette forme d'apprentissage par l'expérience, de pédagogie active, s'étant révélée payante pour nous, nous espérons qu'elle le sera également pour d'autres collègues enseignants et les étudiants. Afin de faciliter l'utilisation de ce livre dans le cadre d'un cours (en licence ou master), l'intégralité des fichiers de données est disponible sur le site de Pearson Education France.

Structure du livre

Le domaine de l'analyse de données étant vaste et complexe, nous avons souhaité organiser ce livre en deux grandes parties distinctes. Une première partie (chapitres 1 à 4) présente les méthodes descriptives en analyse de données (analyses univariées et bivariées, tris croisés, analyses factorielles), la plupart des enquêtes en marketing se limitant aux tests présentés dans cette partie. La seconde partie de l'ouvrage (chapitres 5 à 7) présente un panorama de techniques plus avancées (analyse de variance, régressions, analyse conjointe) afin de guider l'analyste dans ces procédures plus sophistiquées. Enfin, le dernier chapitre traite de la rédaction du rapport, la valeur ajoutée d'une démarche analytique passant aussi par la capacité à communiquer les résultats de manière précise et intelligible.

Remerciements

Nous voudrions remercier vivement ceux qui nous ont aidés à réaliser cet ouvrage, en particulier, Roland Gillet, professeur à l'université Paris 1 Panthéon-Sorbonne et directeur de la collection, pour ses remarques et sa confiance, et Pierre-Louis Dubois, professeur à l'université Paris 2 Panthéon-Assas et à l'ESCP-EAP, pour ses encouragements constants et son aide précieuse. Nos remerciements s'adressent aussi à Laurent Bertrandias (maître de conférences à l'IAE – université Toulouse 1), René Darmon (professeur émérite à l'ESSEC), Laurent Florès (CEO crmmetrix et professeur associé à l'INSEEC), Jean-François Trinquecoste (professeur à l'IAE Bordeaux), Hervé Fenneteau (professeur à l'université Montpellier I), Jean-Philippe Grouthier (administrateur à l'Insee).

Merci aussi à Christophe Lenne et à toute l'équipe de Pearson Education France pour leur travail éditorial constructif et enrichissant.

Enfin, nos plus vifs remerciements vont à nos proches, pour les longs instants volés, le temps étant le plus précieux des cadeaux. Comme le dit Paul Claudel : « *Le temps, tout le consume, et l'amour seul l'emploie.* »

Les auteurs

Manu Carricano est enseignant-chercheur à l'Inseec Paris où il est responsable du département marketing. Il y enseigne le marketing et les études de marchés en licence et master. Il intervient également à l'IAE de Bordeaux dans le master marketing en formation continue. Ses recherches portent sur la convergence des méthodes quantitatives et qualitatives sur Internet ainsi que sur l'optimisation des stratégies de prix. Ses travaux ont fait l'objet de publications et ont été présentés dans des conférences académiques internationales.

Fanny Poujol est titulaire d'un MBA de l'université de Birmingham ainsi que d'un doctorat de l'université Montpellier II. Elle est maître de conférences à l'IAE de Valenciennes, et chercheur associé au laboratoire de recherche Inseec. À l'IAE, elle enseigne la méthodologie, le marketing des services, le commerce international et le management des forces de ventes en licence et master. Elle intervient aussi en master marketing et vente à l'UPMC (Université Pierre-et-Marie-Curie). Ses recherches portent sur la gestion des forces de vente. Ses travaux ont été présentés dans des congrès internationaux (IAE, AFM, EMAC, ANZMAC) et publiés dans des revues académiques (*Décisions Marketing*, *Journal of Business and Industrial Marketing*).

Chapitre 1

Analyser pour décider

Une bonne décision consiste à choisir la plus optimale des solutions parmi une série d'alternatives. Le marketing – et en particulier sa dimension études – s'est longtemps cantonné à un rôle purement descriptif. Mais les bonnes décisions n'arrivent pas par hasard : elles doivent être fondées sur des informations fiables et valides. Tour à tour, les outils d'études de marchés et les techniques d'analyse se sont considérablement enrichis. L'avènement d'Internet, la sophistication et l'exhaustivité des données de panel, la montée en puissance des bases de données clients et du data mining ont repoussé les limites des études de marchés traditionnelles, favorisant l'émergence d'une information marketing de grande qualité et d'analyses explicatives, voire prédictives, des comportements.

Ce chapitre présente les grandes familles d'études de marchés et pose les bases de l'analyse de données en marketing en abordant les concepts de données, de variables et de mesure.

1 Études et recherche en marketing

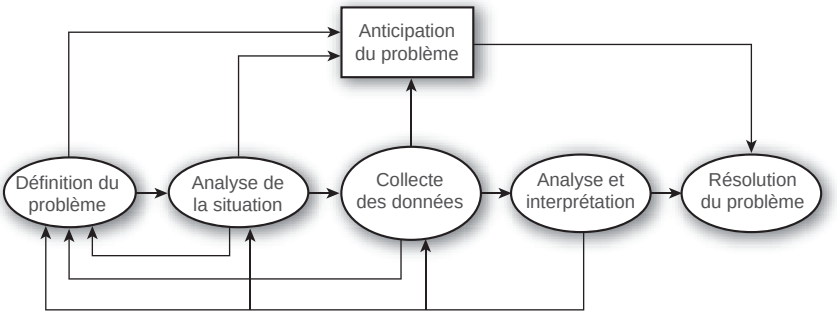
Les études et recherche marketing ont pour but d'aider le responsable marketing à résoudre un problème spécifique, à contrôler ses performances, à planifier les décisions (Evrard, Pras et Roux, 2003). **Leur objectif est de lier l'entreprise à son environnement en développant des instruments de mesure, en collectant et en analysant des données, et en communiquant les résultats et leur interprétation.** Telle est la définition du processus de recherche en marketing qui nous guidera tout au long des huit chapitres de cet ouvrage.

1.1 La démarche d'étude

À partir de la définition précédente, nous pouvons résumer la démarche d'étude à cinq étapes principales, reprises à la figure 1.1 ci-après.

Figure 1.1

Les cinq étapes d'une démarche d'étude.



La première étape de la démarche d'étude est d'identifier le problème managérial : le besoin d'étude est donc déterminé par l'existence d'un problème à résoudre. Plusieurs types de problèmes et plusieurs types de résolutions peuvent être envisagés, comme le montre le tableau 1.1.

Tableau 1.1 : Du problème managérial à la technique d'étude

Problème managérial	Objectifs d'étude	Techniques d'étude
Existe-t-il un marché potentiel pour un nouveau produit ?	– Tester les réactions des consommateurs à l'idée – Tester la composition du produit – Estimer le taux d'essai et de réachat	– Test de concept – Test de formule – Marché-test simulé – Marché-témoin
	– Connaître les concurrents	– Analyse de la concurrence – Panels

Tableau 1.1 : Du problème managérial à la technique d'étude *(suite)*

Problème managérial	Objectifs d'étude	Techniques d'étude
	– Connaître les attentes des consommateurs	– Identifier les bénéfices recherchés par les consommateurs – Étude de segmentation
	– Détecter les forces et faiblesses de la marque	– Étude du capital marque – Analyse des images de marque des concurrents
	– Déterminer un prix de vente	– Test de prix psychologiques – Analyse conjointe

Source : adapté de Vernette, 2000.

La formulation correcte d'un problème permet de faire le lien entre un besoin de décision et la mise en œuvre d'une démarche de recherche, de collecte, d'analyse et d'interprétation d'informations. La fonction « étude » doit donc être envisagée autour de ce paradigme informationnel. Son rôle consiste à transformer des informations brutes en données utiles dans la recherche de nouvelles opportunités, à mettre en place des systèmes d'écoute du marché et de veille concurrentielle, et à prescrire les comportements à adopter sur les marchés. Au confluent des flux d'informations de l'entreprise, elle acquiert aujourd'hui une dimension stratégique croissante.

Si la collecte et l'analyse de données sont au cœur du métier d'analyste en marketing, ces derniers font de plus en plus appel à des données secondaires et à des données stockées dans des entrepôts de données (*data warehouses*). Cette tendance est accentuée par le recours à Internet qui, en combinant habilement sites de marque et techniques de marketing direct, s'avère être une source inépuisable d'informations sur les marchés, les consommateurs, les concurrents.

L'existence de cette profusion de données fait évoluer les besoins d'étude dans l'entreprise et modifie par suite le recours aux différentes techniques. Auparavant, la conduite d'une étude de marché était principalement entendue comme la nécessité de procéder à une collecte de données terrain, souvent par le biais du questionnaire, de l'entretien ou de réunions de consommateurs. Dorénavant, l'accès aisé à des données secondaires, à la dissémination interfonctionnelle d'une intelligence marketing dans l'organisation modifie quelque peu la donne.

Cette vaste quantité d'informations disponibles rend nécessaire, pour le chargé d'étude comme pour le chef de produit, une compétence accrue en analyse de données. Elle permettra d'éviter les erreurs d'interprétation et de maîtriser la qualité d'études souvent réalisées par des instituts. Un besoin croissant d'opérationnalité se fait sentir en la matière. Cette opérationnalité passe tout d'abord par le développement de mesures pertinentes et valides supportant des construits psychologiques (décisions d'achat, notoriété, intérêt pour la marque, le produit, etc.), afin de bien mesurer ce qui se rapporte au problème managérial. Elle passe ensuite par la mise en œuvre d'analyses qui permettent d'expliquer et de prédire des comportements, afin de bien comprendre le problème managérial pour le résoudre et agir.

1.2 Les techniques d'études

Les techniques d'études sont regroupées en deux catégories principales, selon leurs objectifs et leurs limites. Les études quantitatives dominent largement le marché des études, même si, dernièrement, les départements marketing ont manifesté un intérêt croissant pour les études qualitatives. Le tableau 1.2 montre la répartition des différentes techniques en fonction des méthodes de collecte les plus fréquemment utilisées en marketing.

Tableau 1.2 : Répartition des différentes techniques d'étude

Techniques	2005	2006
Quantitatives		
Études quantitatives *via* Internet	13 %	20 %
Études par téléphone	29 %	30 %
Tests en salle	11 %	10 %
Études en face-à-face	42 %	37 %
Études postales	5 %	4 %
Qualitatives		
Réunions de groupe	57 %	55 %
Entretiens individuels	22 %	22 %
Études qualitatives *via* Internet	5 %	17 %
Autres techniques qualitatives	16 %	5 %

Source : adapté de SEMO, 2008 (Syntec Études Marketing et Opinion).

Parmi les techniques les plus largement utilisées, on peut retenir :

- **l'étude *ad hoc*** : étude quantitative ou qualitative réalisée pour le compte d'un seul client ;
- **l'étude omnibus** : étude quantitative réalisée à date régulière. Le questionnaire regroupe l'ensemble des questions de différents souscripteurs ;
- **le baromètre** : étude réalisée à date fixe comme l'omnibus, mais avec le même questionnaire d'une étude à l'autre, pour le compte d'un ou de plusieurs clients ;
- **le panel** : investigation approfondie réalisée périodiquement pour plusieurs clients. Les interviewés sont identiques d'une vague à l'autre. Il s'appuie sur des échantillons importants de 2 000 à 10 000 individus ;

- **le marché-test** : étude quantitative visant à prévoir les ventes et parts de marché d'un nouveau produit ; on parle également de marché-test pour des observations de type expérimental en magasin ;
- **l'entretien individuel** : étude qualitative dont l'objectif est de recueillir le discours individuel. On distingue l'entretien non directif (libre propos), semi-directif (intervention et thèmes), directif (guide d'entretien strict, questions ouvertes), associatif ou projectif (analogie, associations de mots, compléments de phrases, jeux de rôle, etc.) ;
- **la réunion de groupe** : étude qualitative libre et non structurée d'un groupe de 8 à 12 participants, conduite par un animateur. La discussion libre repose sur les phénomènes de psychologie collective des groupes restreints, fondée notamment sur les travaux de Kurt Lewin.

Le tableau 1.3 représente les objectifs et les limites des approches qualitatives et quantitatives.

Tableau 1.3 : Objectifs et limites des approches qualitative et quantitative

Approche	Objectifs	Limites
Qualitative	Répertorier Explorer Générer Comprendre	Généralisation des résultats
Quantitative	Dénombrer Hiérarchiser Pondérer Résumer	Biais déclaratifs Mémorisation des répondants

Source : adapté de Vernette, 2000.

Les études qualitatives sont utilisées dans une dimension principalement exploratoire, afin de comprendre en profondeur des comportements de consommateurs par exemple. Si elles ne permettent pas de généraliser les résultats qu'elles produisent, elles n'en sont pas moins utiles pour dépasser les mesures d'attitudes des questionnaires. Elles permettent d'accéder à une étude approfondie des processus liés aux comportements de consommation, grâce notamment aux entretiens et aux réunions de consommateurs, et d'accéder plus profondément à l'explication de ces comportements, en levant le voile sur des facteurs inconscients (le non-verbal, le « non-dit »), en d'autres termes le monde interne des consommateurs et notamment leur rapport aux marques.

Les études qualitatives se distinguent également par la place qu'elles occupent dans la démarche de recherche. Souvent considérées comme un prélude à l'étude quantitative ou limitées à la confirmation des résultats d'une enquête par questionnaire, elles se substituent de plus en plus aux approches traditionnelles, grâce notamment à l'utilisation d'Internet et de ses potentialités multimédias, et à la nécessité croissante pour le marketing d'être connecté au terrain. Garnier, par exemple, a lancé, il y a peu, une vaste opération de type ethnographique baptisée *Consumer Connect*, dont l'objectif était avant tout d'immerger les chefs de produit parmi les consommateurs et d'observer leur utilisation du produit *in situ*. L'avènement d'Internet a contribué à repopulariser cette technique auprès des instituts d'étude : on peut citer l'émergence de la netnographie (voir ci-après) ou encore le *Home Use Blog* (HUB), développé conjointement par Danone et la société Repères.

EXEMPLE

La netnographie

On constate, depuis quelques années, un intérêt grandissant pour l'information collectée à partir de l'observation de communautés virtuelles, nouvelles formes de communautés dont Internet a permis l'émergence. Ainsi, de nombreuses firmes ont réalisé des études sur la base d'informations issues de forums de discussion et n'ont pas tardé à saisir les opportunités offertes par ces nouveaux types d'interactions sociales. Kozinets a développé récemment une approche nouvelle – l'ethnographie sur Internet ou netnographie – qu'il définit comme « une nouvelle méthode de recherche qualitative qui adapte la méthode de l'ethnographie à l'étude des cultures et des communautés qui émergent grâce aux communications informatisées » (Kozinets, 2002, p. 62). En tant que technique de recherche en marketing, la netnographie utilise l'information publique disponible sur les forums en ligne afin d'identifier et de comprendre les besoins et les influences qui pèsent sur les décisions d'achat de groupes de consommateurs présents sur Internet. Pour Laurent Florès, CEO de la société d'étude crmmetrix, spécialiste de l'écoute client, le canal Internet permet aux marques de participer à de véritables conversations et de s'appuyer sur un puissant levier du marketing : le bouche à oreille. Il est désormais possible de quantifier le volume de ces conversations, d'analyser leur contenu et le profil des intervenants, avec un avantage important sur les techniques traditionnelles, puisque cette approche n'altère pas le contexte étudié par l'intervention d'un analyste mais collecte plutôt une information en langage naturel.

Les techniques quantitatives, auxquelles cet ouvrage est essentiellement consacré, constituent la part dominante des études marketing. Leur objectif est avant tout de mesurer, de quantifier et de permettre de généraliser les résultats à partir de l'échantillon de la population concernée. Ce type d'étude repose généralement sur un grand nombre d'observations et sur des informations structurées (valeurs numériques, échelles ou valeurs nominales) par opposition aux informations non structurées (discours, texte libre/questions ouvertes, etc.). Plus précisément, trois types d'études quantitatives peuvent être distingués, en fonction du contexte de découverte de l'information : décrire, expliquer, prédire.

Les **études descriptives** sont fondées sur des mesures dont le but est de collecter des données brutes afin de créer des structures décrivant les caractéristiques d'une population cible ou d'un marché. Elles peuvent être utiles, entre autres, pour faire la photographie d'un marché, de la satisfaction des consommateurs, de la notoriété d'une marque. La dimension descriptive est l'objectif premier traditionnellement assigné aux études marketing. Cette étape importante a pour objet de mesurer la force d'association entre deux variables, par exemple, et permet de poser un cadre d'analyse nécessaire aux études explicatives et prédictives.

Les **études explicatives** ont pour objet de transformer des données brutes en structures expliquant des relations de causalité entre deux ou plusieurs variables. L'approche explicative est utile lorsque l'étude a pour objectif de comprendre les causes directes d'un phénomène. Ce type d'étude peut permettre, par exemple, de modéliser l'impact de la publicité sur les ventes. L'approche explicative est particulièrement utile dans un contexte d'aide à la décision, où le but assigné à l'étude n'est plus simplement de décrire mais aussi de comprendre, de la manière la plus fiable et la plus valide, les déterminants affectant la performance des décisions marketing.

Les **études prédictives**, quant à elles, ont pour objet de transformer les données brutes collectées sur les caractéristiques comportementales des consommateurs ou des entreprises/marchés pour créer des modèles prédictifs à des fins d'optimisation. Ces approches,

surtout utilisées dans des contextes de gestion de la relation client, nécessitent des observations en très grand nombre et des outils sophistiqués (voir focus 1.1). Pour notre part, dans les chapitres suivants, nous nous concentrerons principalement sur les deux premiers types d'étude.

Focus 1.1 Le data mining

Le data mining, ou fouille de données, est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de bases de données informatiques (souvent de grande taille), de façon automatique ou semi-automatique, en vue de détecter des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données. En bref, le data mining est l'art d'extraire des informations, voire des connaissances à partir de données. Le data mining est soit descriptif, soit prédictif : les techniques descriptives en data mining visent à mettre en évidence des informations présentes mais cachées par le volume des données (c'est le cas des classifications automatiques d'individus et des recherches d'associations de produits) ; les techniques prédictives visent à extrapoler de nouvelles informations à partir des informations présentes, ces nouvelles informations pouvant prendre la forme de classements ou scorings (sélection de clients selon certains critères), ou de prédictions comme l'appétence pour un produit (probabilité d'achat futur) ou le risque d'attrition (probabilité de départ à la concurrence).

Source : adapté de Tufféry, 2005.

2 Des données aux variables

La plupart des entreprises sont aujourd'hui noyées sous l'information mais elles ont soif de connaissance. C'est la capacité de l'analyste à mettre en œuvre une démarche analytique qui permet de créer, de gérer et de diffuser cette connaissance dans l'organisation. Ce processus repose sur trois concepts que nous allons maintenant définir : les données, l'échantillon et les variables.

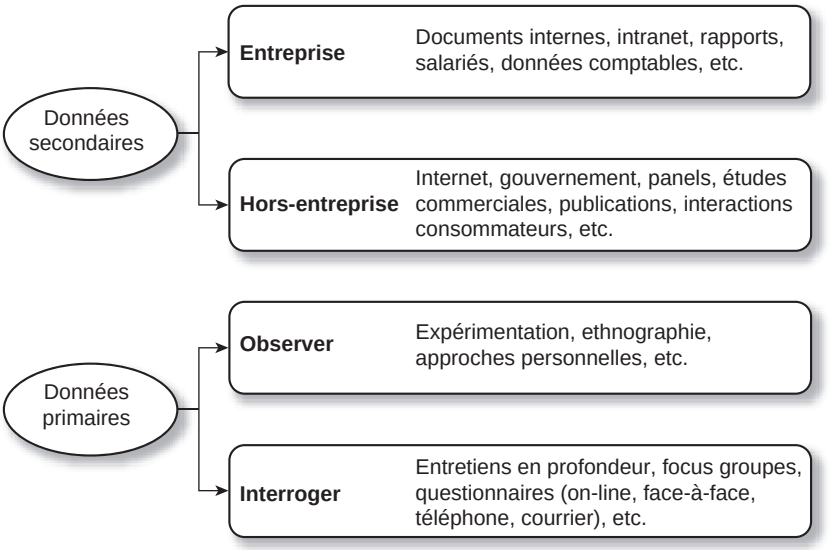
2.1 Types de données

Les types de données en marketing sont en général identifiés en fonction de leur source. Ainsi, on distingue les données secondaires et les données primaires (voir figure 1.2).

Les **données secondaires** sont des données qui ont été collectées préalablement à l'étude, pour répondre à d'autres problèmes, ce qui peut fortement en limiter la pertinence et la précision (Malhotra *et al.*, 2007). Elles sont cependant d'un accès facile et leur coût est relativement faible. Souvent perçues comme des données à faible valeur ajoutée en marketing (limitées à une définition de problème, voire à quelques tests pour mettre en valeur les résultats principaux), elles occupent désormais une place de plus en plus importante dans les études. Les sources d'information de cette nature sont aujourd'hui abondantes et doivent systématiquement être prises en considération avant toute collecte de données primaires. On distingue les **données secondaires internes**, issues de l'entreprise (reportings, intranet, données comptables, informations émanant des salariés…) et les **données secondaires externes**, issues de l'environnement de l'entreprise :

Figure 1.2

Les sources de données.



- en premier lieu Internet qui donne accès à des données structurées (fichiers logs, cookies, etc.) et surtout à des données non structurées (blogs, forums, interactions sociales, etc.) que l'on commence aujourd'hui à analyser, à traiter (netnographie, analyse lexicométrique, text mining, etc.) ;
- les données gouvernementales : données de recensement, données macroéconomiques, etc. ;
- les données de panels (consommateurs, distributeurs, audience) ;
- les études de marchés publiées de nature commerciale ;
- les interactions avec les consommateurs ou points de contacts : lettres de réclamations, call-centers, e-mails reçus, etc.

Focus 1.2 Les panels

Les panels ont considérablement évolué ces dernières années : gains de restitution de l'information, amélioration de la couverture des circuits de distribution (données de panels et données issues du scanning en sortie de caisse), offre enrichie (information accessible en ligne, analyses spécifiques des variables du mix et de leur performance). On distingue, en France, plusieurs types de panels largement plébiscités (42 % du marché des études) : les panels de consommation, les access panels (ou panels de consommateurs), les access panels on-line, les panels de distributeurs (ou panels de détaillants) et les panels d'audience.

- **Panels de consommation** : recueil d'informations sur leurs achats auprès d'un échantillon de ménages. Ce type de panel permet de répondre aux questions : « Qui consomme quoi ? » et « En quelle quantité ? » (taux de pénétration, quantités achetées, etc.) et de mesurer l'évolution de la consommation dans le temps.
- **Access panels** : recueil d'informations auprès d'individus ou de foyers représentatifs de la population nationale, qui ont accepté de participer à des enquêtes ponctuelles. Ils sont interrogés sur leurs pratiques, leurs opinions, leurs goûts et leurs préférences, pour des études *ad hoc* (tests de produits, de concepts, études d'usages et d'attitudes, tracking, etc.).
- **Access panels on-line** : recueil d'informations auprès d'internautes panélisés qui ont accepté de participer à des enquêtes ponctuelles. Le recrutement se fait le plus souvent *via* un site de recrutement sur Internet. Ils sont interrogés sur leurs pratiques, leurs opinions, leurs goûts, leurs préférences.

- **Panels de distributeurs** : recueil d'informations auprès d'un échantillon de points de vente afin de connaître les volumes, les prix de vente, les parts de marché de différentes marques d'un segment, d'évaluer la présence de la marque dans les différents canaux de distribution (distribution numérique, distribution valeur), de suivre les évolutions de la distribution, l'offre disponible dans les points de vente (linéaires accordés aux différentes marques, ruptures, promotions), de mesurer l'impact sur les ventes d'une modification de l'offre (promotion, lancement, etc.). *Infoscan Census* (panel d'Information Resources Inc., IRI) est le premier du genre à abandonner la méthode de l'échantillon au profit d'une remontée exhaustive des magasins.
- **Panels d'audience** : échantillon représentatif de foyers dont on mesure l'écoute des différentes chaînes de télévision. Il n'existe qu'un seul panel depuis l'arrêt du panel Sofres-Nielsen : *Mediamat*, de Médiamétrie, panel de 3 100 foyers, soit 8 000 individus de 4 ans et plus équipés d'un audimètre.

Les **données primaires** sont des données qui ont été collectées dans le but de résoudre le problème managérial propre à l'étude. Il s'agit de données brutes, qui doivent être préparées, analysées puis interprétées (Hair *et al.*, 2006). Dans ce cas, les cinq étapes de la démarche d'étude doivent être respectées. Ce chapitre étant consacré à l'étape de la collecte des données, les chapitres suivants aborderont l'analyse et l'interprétation des résultats pour une série de tests pouvant être mis en œuvre dans une démarche d'étude ou de recherche marketing.

2.2 L'ÉCHANTILLON

Afin de bien illustrer les étapes d'un plan de sondage, un petit détour historique peut s'avérer intéressant. Tout commence aux États-Unis, lorsque Franklin D. Roosevelt se représente contre Alf Landon aux élections de 1936. Derrière les candidats, deux hommes s'affrontent pour pronostiquer le résultat de ces élections. D'une part Codely, rédacteur en chef du *Literary Digest*, utilise la technique du vote de paille (*straw vote*) : quelques jours avant les élections, il fait paraître des bulletins de vote dans son journal et demande à ses lecteurs de mentionner leur choix. Il reçoit 2,4 millions de réponses et donne Landon gagnant. D'autre part, Gallup, créateur de l'institut éponyme, n'interroge que 4 000 personnes et joue Roosevelt gagnant. La victoire de ce dernier marque la naissance des instituts de sondage. Gallup est le père de l'échantillon représentatif, le premier à avoir eu l'idée de reconstituer une population en miniature. Deux ans après, les sondages sont importés en France par Jean Stoetzel, philosophe et sociologue, créateur en 1938 de l'Institut français d'opinion publique (IFOP). Cette jeune pratique est construite autour de deux étapes principales : la définition de la population à étudier et la sélection de l'échantillon.

La population à étudier doit être définie avec le plus grand soin (par exemple les clients d'une enseigne de distribution). Cette définition inclut celle des unités de sondage (l'individu détenant l'information) qui sont l'objet de l'observation. Dans de nombreux cas, en marketing, on ne se préoccupe pas de l'ensemble de la population mais plutôt des consommateurs de tel ou tel produit, ou catégorie de produits, qui constituent la cible des actions envisagées.

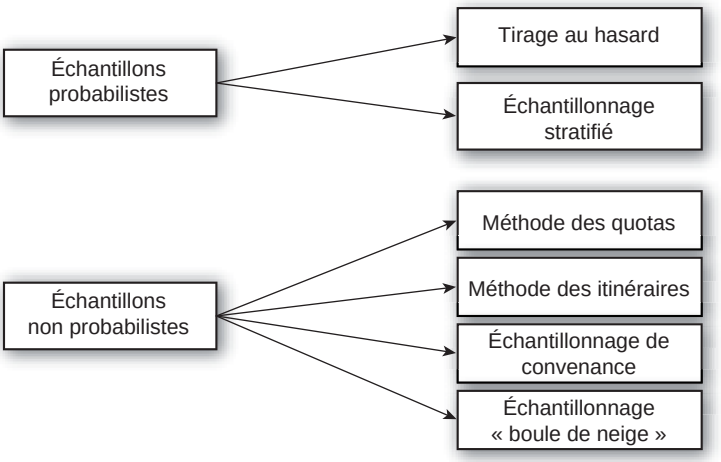
Vient ensuite l'étape du choix de l'échantillon et de sa taille. Deux méthodes principales sont utilisées, dont l'objectif est de sélectionner un échantillon assurant la meilleure précision possible des résultats au moindre coût (pour une description détaillée, voir Evrard *et al.*, 2003). La **méthode probabiliste**, dans laquelle chaque individu de la population

concernée a une probabilité connue d'appartenir à l'échantillon, permet d'obtenir des échantillons représentatifs. Généralement utilisée sur de grands échantillons, elle s'avère très coûteuse (l'Insee, par exemple, la pratique en France). Les **méthodes non probabilistes** (ou méthodes empiriques) permettent de constituer un échantillon résultant d'un choix raisonné qui vise à le faire ressembler à la population dont il est issu. Parmi ces méthodes, les instituts de sondages français recourent volontiers à la méthode dite des quotas, qui, bien que scientifiquement moins précise, moins fiable que la méthode aléatoire, présente l'énorme avantage de pouvoir s'appliquer à des échantillons plus réduits, de coûter moins cher et d'être mise en œuvre beaucoup plus rapidement. Ainsi, un sondage au téléphone selon la méthode des quotas peut être réalisé en moins de 48 h.

La figure 1.3 détaille les méthodes d'échantillonnage qui seront définies dans la section suivante.

Figure 1.3

Les méthodes d'échantillonnage.



- **Tirage au hasard** : l'échantillon aléatoire consiste à tirer au hasard un individu de la population avec une probabilité connue et différente de zéro d'appartenir à l'échantillon. La connaissance de cette probabilité d'appartenance de l'individu à l'échantillon permet de calculer la marge d'erreur sur les résultats obtenus (voir focus 1.3). Ce type d'échantillonnage permet de protéger les utilisateurs des résultats contre une sélection biaisée de l'échantillon (même si les risques de biais dus au questionnaire et aux non-réponses subsistent). On parlera de tirage aléatoire simple si les individus qui composent la population ne font l'objet d'aucun regroupement avant tirage.
- **Échantillonnage stratifié** : dans le cas où les variables étudiées sont fortement dispersées, c'est-à-dire dans ceux où des classes seraient sous- ou surreprésentées en raison du tirage au hasard, il peut s'avérer utile d'utiliser des variables dites de stratification, qui permettent de réaliser une répartition de la population en classes appelées « strates ». La stratification sera d'autant plus efficace pour améliorer la précision que les strates seront homogènes par rapport aux variables étudiées. Dans le cas d'une étude sur des points de vente, cette variable pourra être le fait d'être client ou non.
- **Méthode des quotas** : cette méthode, la plus utilisée en France, reprend les principes de qualification de l'échantillonnage stratifié. Elle est moins coûteuse que les méthodes aléatoires où l'enquêteur, en cas d'absence de la population, doit revenir/rappeler jusqu'à 3 ou

4 fois à l'adresse/au numéro qui lui a été indiqué. Cette méthode présente l'avantage de la simplicité : on choisit quelques caractéristiques dont on connaît la distribution statistique dans la population étudiée (par exemple, sexe, âge, catégorie socioprofessionnelle [CSP] du chef de famille), puis on donne à chaque enquêteur un plan de travail qui lui impose le respect de certaines proportions au sein des interviewés. Cette méthode, par opposition aux deux méthodes précédentes, donne des estimations biaisées car les différentes catégories de population présentent des probabilitiés différentes et inconnues d'être touchées par un enquêteur. D'autre part, la méthode des quotas ne permet théoriquement pas de calculer les marges d'erreur associées aux résultats trouvés, comme une méthode aléatoire permet de le faire.

- **Méthode des itinéraires** (ou *random route*) : dans une commune, par exemple, on impose à l'enquêteur un point de départ et un itinéraire à suivre, avec tirage systématique des logements dans lesquels il doit effectuer des interviews (par exemple, interroger les foyers toutes les trois portes dans un immeuble).
- **Échantillonnage de convenance** : il est conçu par l'enquêteur pour des raisons de praticité. Il fait généralement appel à des personnes interceptées dans la rue, à la sortie des caisses en magasin, etc. C'est la moins coûteuse et la plus rapide de toutes les techniques d'échantillonnage mais elle présente de fortes limites : biais de sélection, non-représentativité. Il n'est donc théoriquement;pas significatif de généraliser les résultats.
- **Échantillonnage « boule de neige »** : on choisit un premier groupe de répondants, au hasard généralement, puis on leur demande d'indiquer d'autres répondants potentiels appartenant à la population ciblée. Cette méthode peut être utile pour des enquêtes sur les leaders d'opinion par exemple.

La détermination de la **taille de l'échantillon** est une étape cruciale en analyse de données. Un échantillon trop petit peut induire une perte d'informations importante ou empêcher la réalisation de nombreux tests soumis à des contraintes en termes de nombre d'observations. À l'inverse, un échantillon trop important constitue une perte de temps et de budget dommageable pour la réussite de l'étude. Il est important de noter que la précision de l'information recueillie dépend principalement de la taille de l'échantillon et non du **taux de sondage** (défini par le rapport n/N, où *n* est la taille de l'échantillon et *N* celle de la population).

Dans la pratique, les chargés d'études utilisent des abaques (feuilles de calcul) donnant la taille de l'échantillon en fonction du degré de précision des résultats que l'on veut obtenir. Certains professionnels des études considèrent qu'il n'y a pas de raison valable de travailler avec des échantillons de plus de 1 000 à 1 500 répondants. En effet, s'il est admis que la précision des résultats est influencée par la taille de l'échantillon, cette influence s'avère négligeable au-dessus de 1 500 observations. Pour trouver la taille adéquate de l'échantillon N, une règle empirique facile à appliquer – même si elle est contestable d'un point de vue purement statistique – consiste à partir de l'erreur, exprimée en pourcentage, que l'on est prêt à tolérer :

$$N = 1/\text{erreur}^2.$$

Par exemple, si l'on accepte une erreur de +/–5 % au niveau de la précision des résultats, on obtient une taille d'échantillon de $1/0{,}05^2$, soit 400 répondants.

Les éléments liés à la précision de la mesure sont centraux en analyse de données. Nous détaillons la méthode de calcul de l'intervalle de confiance dans le focus 1.3. Les éléments de discussion dépendant du principe de test statistique seront abordés dans le chapitre 2.

Focus 1.3 L'intervalle de confiance

La précision statistique d'un test (proportion ou moyenne) s'exprime en calculant l'intervalle de confiance, qui indique la marge d'erreur lorsqu'on généralise une estimation obtenue sur un échantillon à l'ensemble de la population représentée. La longueur de l'intervalle diminue lorsque la taille de l'échantillon augmente.

On retient la formule suivante pour calculer l'intervalle de confiance d'une proportion :

$$p - z\sqrt{\frac{pq}{n}} \leq \pi \leq p + z\sqrt{\frac{pq}{n}}$$

où :

p = pourcentage observé dans l'échantillon ;

q = 1 – p ;

z = valeur dérivée de la loi normale centrée réduite, égale à 1,96 si $\alpha = 0{,}05$ (degré de confiance) ;

π = pourcentage réel dans la population mère ;

n = taille de l'échantillon.

L'intervalle de confiance d'une moyenne *m* sur *n* individus avec un écart type σ se calcule de la manière suivante :

$$m - z\frac{\sigma}{\sqrt{n}} \leq \pi \leq m + z\frac{\sigma}{\sqrt{n}}$$

Prenons l'exemple suivant : un sondeur réalise une étude d'audience par téléphone pour connaître les caractéristiques sociodémographiques et les comportements – notamment en termes de dépenses en SMS – des téléspectateurs de la *Super Star,* émission de télé-réalité diffusée en prime time sur le câble et le satellite. Il sélectionne 1 000 numéros de téléphone par tirage aléatoire simple dans la base de données des abonnés de la chaîne (qui en compte 120 000 sur le câble et 2 100 000 sur le satellite). On pose l'hypothèse que les 1 000 personnes répondent effectivement aux enquêteurs. On constate que l'émission absorbe 36,8 % de l'audience des personnes interrogées de moins de 35 ans, et que le montant moyen dépensé par cette cible en SMS et appels téléphoniques est de 6,2 €, avec un écart type de 2,2 €.

Le montant moyen dépensé par ces abonnés est de :

$$6{,}2 - 1{,}96\frac{2{,}2}{\sqrt{1000}} \leq \pi \leq 6{,}2 + 1{,}96\frac{2{,}2}{\sqrt{1000}}$$

Soit : $6{,}06 \leq \pi \leq 6{,}33$

L'audience moyenne des abonnés de moins de 35 ans est de :

$$0{,}368 - 1{,}96\sqrt{\frac{(0{,}368 * 0{,}632)}{1000}} \leq \pi \leq 0{,}368 + 1{,}96\sqrt{\frac{(0{,}368 * 0{,}632)}{1000}}$$

Soit : 33,6 % $\leq \pi \leq$ 39,7 %

Le sondage réalisé permet donc d'estimer cette proportion avec une précision absolue de 3,2 % (au degré de confiance 0,95).

SPSS

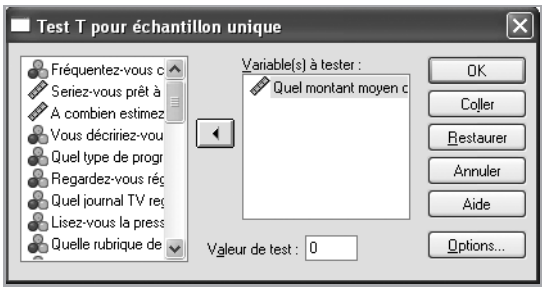
SPSS permet également d'estimer l'intervalle de confiance d'une mesure. L'exemple retenu ici servira de fil rouge tout au long de cet ouvrage. Une enseigne de grands magasins souhaite ouvrir un nouveau point de vente, mais elle ne le fera que si le potentiel de marché est suffisant. Une enquête a donc été réalisée sur 400 répondants, en face-à-face. Parmi les questions posées, les enquêteurs ont relevé l'intérêt des répondants pour l'ouverture du nouveau point de vente, ainsi que le montant qu'ils seraient prêts à dépenser.

Il est possible d'identifier l'intervalle de confiance d'une moyenne avec SPSS en utilisant la procédure du test *t* pour échantillon unique.

Ouvrez le fichier exemple « pointdevente.sav » disponible comme tous les fichiers d'exercices sur le site de l'ouvrage [1]. Allez dans le menu **Analyse > Comparer les moyennes > Test T pour échantillon unique...** Une boîte de dialogue s'affiche (voir figure 1.4).

Figure 1.4

Test t pour échantillon unique sous SPSS.



Faites glisser dans la boîte de dialogue la variable à tester *montant* qui correspond à la question suivante : « Quel montant moyen dépensez-vous par mois dans ce type de point de vente ? »

L'analyse donne les résultats présentés à la figure 1.5.

Figure 1.5

Statistique sur échantillon unique.

Statistiques sur échantillon unique

	N	Moyenne	Ecart-type	Erreur standard moyenne
Quel montant moyen dépensez-vous par mois dans ce type de point de vente?	400	153.5100	91.14782	4.55739

Le premier résultat donne le nombre de répondants, la moyenne, l'écart type et l'erreur standard.

Le second résultat donne l'intervalle de confiance pour un degré de confiance de 95 % qui se situe entre 144,55 € et 162,46 € (voir figure 1.6). Les éléments d'interprétation liés à la théorie des tests statistiques seront approfondis au chapitre 2.

Figure 1.6

Test sur échantillon unique.

Test sur échantillon unique

	Valeur du test = 0					
	t	ddl	Sig. (bilatérale)	Différence moyenne	Intervalle de confiance 95% de la différence	
					Inférieure	Supérieure
Quel montant moyen dépensez-vous par mois dans ce type de point de vente?	33,684	399	,000	153.51000	144.5505	162.4695

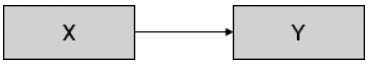
1. Vous trouverez ce fichier à l'adresse : http://www.pearsoneducation.fr.

Notons également qu'un des modules de SPSS (*SamplePower*) permet d'optimiser la combinaison entre la puissance du test, l'intervalle de confiance et la taille de l'échantillon. L'approche est fondée sur des tests de moyenne et de différences de moyennes, des tests de proportions et de différences de proportions, des analyses de variance, entre autres.

2.3 La notion de variable

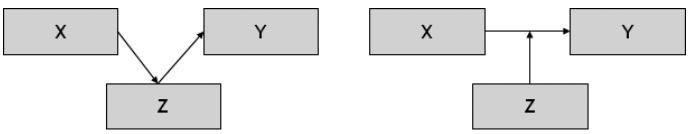
Le principe de modélisation, sous-jacent à l'analyse des données, impose de définir la notion de variable. La modélisation est entendue ici comme la réalisation d'une représentation simplifiée d'un phénomène, la variable étant l'expression du caractère observé dans la population. La formulation la plus simple d'un modèle vise à définir une relation de cause à effet entre deux natures de variables : les variables indépendantes (ou variables explicatives) et les variables dépendantes (ou expliquées). Dans ce modèle simple, la variable indépendante représente la cause, dont l'effet se mesure sur la variable dépendante (voir figure 1.7). Ce modèle permet, par exemple, de représenter le lien entre la fréquence d'achat et la fidélité au point de vente.

Figure 1.7
Relation causale simple.



D'autres variables peuvent intervenir dans cette relation directe entre la (ou les) variable(s) indépendante(s) et la (ou les) variable(s) dépendante(s) [Thiétart *et al.* 1999, p. 339]. Dans le premier cas, l'effet de la variable indépendante X sur la variable dépendante Y se mesure par l'intermédiaire d'une troisième variable dite « médiatrice ». L'association ou la causalité observée entre X et Y résulte du fait que X influence Z qui à son tour influence Y. Cette variable Z – le montant dépensé par exemple – peut intervenir dans la relation entre la fréquence d'achat et la fidélité au magasin. Dans le second cas, la présence de la variable modératrice modifie l'intensité (c'est-à-dire l'amplifie ou la diminue) et/ou le signe de la relation entre la variable indépendante et la variable dépendante. On pourra mesurer l'effet de cette variable modératrice par exemple en décomposant la population en sous-populations (classes d'âge, segments de clients, etc.) et en testant la relation dans les sous-groupes pour y vérifier le type d'effet (voir figure 1.8).

Figure 1.8
Effets médiateur et modérateur.



Les tests statistiques mis en œuvre pour mesurer ces relations seront sélectionnés en fonction de l'objectif de l'enquête (voir section 3 : Mesurer à l'aide d'un questionnaire) et en fonction des variables collectées. Les variables sont de deux types :

- **qualitatives** : leurs modalités, c'est-à-dire la manière dont les observations sont regroupées, ne peuvent être calculées ;

- **quantitatives** : leurs modalités sont mesurables et les tests envisageables sont nombreux.

Le marketing et plus largement les sciences sociales s'intéressent également à la mesure de phénomènes mentaux, telles les opinions, les attitudes ou encore les préférences, au travers d'indicateurs : les **échelles de mesure**. Ces dernières ont pour objet de fournir au répondant un support d'expression de ces phénomènes complexes à observer, le plus souvent sous la forme d'échelles de notation :

- L'**échelle nominale** a pour principe d'utiliser les nombres comme des étiquettes afin de classer chacune des modalités. Les échelles nominales peuvent être utilisées pour identifier des classes d'individus. Par exemple, on peut utiliser la nomenclature des catégories socioprofessionnelles (CSP) ou encore identifier des marques lors d'une étude de notoriété assistée, identifier des attributs de produits. Dans l'échelle nominale, chacune des modalités de la variable est équivalente aux autres.

  Exemple :

  Êtes-vous ? 1. ☐ Un homme 2. ☐ Une femme

- L'**échelle ordinale** est une échelle de classement comme l'échelle nominale, dans laquelle les nombres attribués à chaque modalité ont une relation d'ordre avec un continuum sous-jacent. On peut, par exemple, utiliser une échelle ordinale pour classer des préférences de marques. L'échelle ordinale permet en effet de déterminer les relations d'ordre en calculant les fractiles, les déciles et les médianes à partir de ces données (voir chapitre 2).

  Exemple :

  Notez de 1 à 5 la qualité gustative du produit X
  (1 étant la note la plus faible, 5 la note la plus élevée) :

1	2	3	4	5

- L'**échelle métrique** possède les propriétés des échelles nominale et ordinale, mais elle permet également de comparer les distances entre les objets, les modalités étant séparées par des espaces équidistants. L'exemple le plus courant est celui du thermomètre, la différence entre 0 °C et 1 °C étant la même qu'entre 1 °C et 2 °C, etc. L'échelle métrique est la plus couramment utilisée en marketing, même si pour ces mesures d'attitudes les intervalles ne sont pas toujours équidistants. Appartiennent à cette catégorie, l'échelle de différentiel sémantique d'Osgood ou l'échelle de Stapel, qui ont pour but de conduire à l'élaboration de profils de répondants, l'échelle d'intensité de Likert ou échelle d'accord, les échelles d'intention.

  Exemples :

  Échelle d'Osgood

  Avez-vous trouvé que le goût du produit X était ?

Mauvais	1	2	3	4	5	Bon

Échelle de Stapel

Choisissez un nombre positif si vous pensez que le mot décrit bien le produit X, un nombre négatif si vous pensez que le mot ne décrit pas bien le produit X, en notant de +5 à –5 :

Bon	….
Utile	….
Pratique	….
etc.	

Échelle de Likert

(Pas du tout d'accord)	1	2	3	4	5	(Tout à fait d'accord)

Échelle d'intention

Si la marque M lançait ce type de produit :

Je n'achèterai certainement pas ce produit	1	2	3	4	5	J'achèterai certainement ce produit

SPSS

Dans SPSS, ces trois types de variables (nominale, ordinale et métrique) doivent être définis dans la partie **Affichage des variables** de l'éditeur de données (lorsque le fichier a été ouvert), dans la colonne **Mesure**.

3 Mesurer à l'aide d'un questionnaire

La construction d'un questionnaire amène à s'interroger sur la mesure des concepts. Comment mesurer, par exemple, la qualité du service ? Même en cas d'études se fondant sur une seule question pour mesurer un concept, il est recommandé d'avoir recours à des échelles de mesure. L'objectif d'une échelle est d'éviter d'avoir à représenter un phénomène abstrait – un construit – par le biais d'une seule et unique variable, en privilégiant l'utilisation d'indicateurs qui permettent de représenter les différentes facettes de ce construit. Ainsi, un chargé d'étude qui chercherait à mesurer la satisfaction vis-à-vis d'une marque pourrait poser une question unique : « Êtes-vous satisfait ? » et fonder son analyse sur cette seule réponse. De manière évidente, le fait de développer une mesure de la satisfaction à partir d'un ensemble d'items (de libellés) dont on sait (par des études préalables ou par le biais de la théorie) qu'ils mesurent correctement la satisfaction, permet de collecter des réponses mieux orientées et d'estimer la fiabilité de la mesure effectuée, non plus à partir d'une réponse mais plutôt à partir d'une forme de réponse « *moyenne* » à une série de questions associées. On mesurera donc la satisfaction en interrogeant des clients sur la satisfaction globale, la propension à recommander le produit et la probabilité de réachat par exemple.

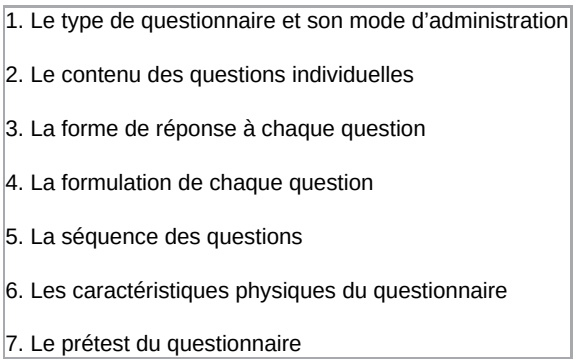
Ainsi, il est généralement recommandé d'utiliser plusieurs items pour mesurer un concept et donc de commencer par chercher s'il existe un outil de mesure du concept que l'on souhaite évaluer. S'il n'existe pas d'échelle (parce que le concept est nouveau ou qu'il s'applique à un domaine particulier), il faut en créer une.

3.1 Le développement des outils de mesure

Churchill, qui est souvent pris comme référence dans la construction d'échelles de mesure, précise que, pour remplir son véritable rôle, « le questionnaire doit susciter et maintenir l'intérêt du répondant » (Churchill, 1998, p. 343). Pour ce faire, le chargé d'étude doit se poser un certain nombre de questions afin de limiter divers biais. Ces questions sont présentées à la figure 1.9.

Figure 1.9

Les étapes du développement du questionnaire.



La forme du questionnaire et son mode d'administration ne sont pas neutres. L'introduction du questionnaire doit présenter les objectifs de l'étude et préciser le caractère confidentiel de l'enquête. Une phrase d'accroche peut rassurer le répondant et l'inciter à répondre de manière authentique.

Le choix du contenu des questions est une étape fondamentale. Il est recommandé de définir le cadre conceptuel dans lequel se situent l'étude et les concepts de base, à l'aide d'une revue de la littérature. L'étude des articles académiques et des ouvrages déjà parus sur le sujet permet d'approfondir la définition du concept étudié. Cette étape permet aussi de trouver des instruments de mesure préexistants qui peuvent être réutilisés, traduits ou encore adaptés.

On utilise la plupart du temps des échelles de mesure préexistantes. Afin de valider dans un contexte français ces instruments de mesure (souvent anglo-saxons), un certain nombre d'étapes doivent être respectées (voir focus 1.4).

Focus 1.4 La traduction d'échelles de mesure

Il existe différentes méthodes pour traduire un questionnaire :

- la **méthode traditionnelle** : le chercheur effectue seul la traduction ou avec l'aide de traducteurs professionnels. Cette méthode est peu recommandée si le chercheur n'est pas parfaitement bilingue ou si les traducteurs sont extérieurs au domaine de la recherche ;
- la **méthode du comité** : le chercheur réunit un ensemble de chercheurs qui maîtrisent parfaitement la langue et qui sont spécialistes de son domaine de recherche. Tous les items sont traduits. Le problème de cette méthode est qu'il est difficile de réunir un tel comité d'experts ;

- la **rétro-traduction** : des chercheurs bilingues sont sollicités pour traduire le questionnaire. Les traductions sont ensuite remises à des chercheurs dont la langue d'origine est celle du questionnaire, afin qu'ils le traduisent. Cette dernière version est ensuite comparée avec la version originale pour repérer les éventuelles différences.

Lorsque le chargé d'étude ne trouve pas d'échelle et souhaite développer son propre outil de mesure, il est préconisé de réaliser une étude exploratoire (entretiens, méthode des incidents critiques…). Par exemple, une recherche sur la satisfaction au travail des commerciaux fait ressortir plusieurs composantes : les relations avec les pairs, avec la hiérarchie, avec les clients, etc. Une étude qualitative a ainsi été réalisée auprès d'un échantillon de 30 vendeurs, auxquels on demandait ce qui les satisfaisait dans leur travail.

À partir de la définition retenue, on dresse ensuite une liste d'items à partir de la littérature (items préexistants empruntés à d'autres échelles ou adaptés) ou à partir d'une étude de terrain exploratoire (extraction de verbatims [1]). Le pool d'items est ensuite soumis à un ou deux panels d'experts (chercheurs ou praticiens du domaine) qui éliminent les énoncés ne leur paraissant pas adéquats à la mesure du concept. Il s'agit ici d'évaluer ce que l'on appelle la « validité faciale du concept » (voir chapitre 3).

Nous avons présenté au point 2.3 les différentes formes possibles d'échelles. Dans un souci de neutralité et de symétrie, il faut veiller à ce que l'équilibre des réponses positives *vs* négatives autour du point médian placé au centre soit respecté. Le choix du nombre d'alternatives de réponse se fait par arbitrage : l'attention des répondants faiblit aussi avec le nombre de questions et de modalités de réponses.

En ce qui concerne la formulation et la séquence des questions, il est généralement recommandé d'alterner l'ordre des questions et le sens des interrogations, afin de limiter les effets de lassitude, de halo ou de contamination. L'effet de halo se manifeste lorsqu'une suite de questions est posée dans le même sens : la personne interrogée peut alors avoir tendance à répondre toujours de la même manière alors que l'effet de contamination concerne l'influence directe d'une question sur les questions suivantes.

Enfin, il est très important de tester le questionnaire avant de l'administrer, d'une part, pour vérifier que toutes les questions sont bien comprises et qu'elles n'engendrent pas de blocage et, d'autre part, afin de tester le temps nécessaire pour y répondre. Une vingtaine de répondants peuvent s'avérer nécessaires pour effectuer ce type de prétest de compréhension.

3.2 Le concept de mesure

L'acte de mesurer est l'opération par laquelle on fait correspondre à une donnée collectée une grandeur considérée comme capable de représenter le phénomène décrit par la donnée. Cette opération est affectée par un certain nombre d'éléments qui éloignent la mesure idéale de la mesure obtenue. Le modèle de la vraie valeur (Evrard *et al.*, 1997, p. 287) consiste à décomposer le résultat d'une mesure en ses différents éléments : la vraie valeur (censée représenter la mesure parfaite) et les termes d'erreur (erreur aléatoire et erreur systématique), comme le montre l'équation suivante :

1. L'extraction de verbatims fait partie des stratégies d'analyse d'un corpus textuel (type retranscriptions d'entretiens). Les verbatims permettent de nourrir l'analyse de citations des personnes interrogées.

M (mesure obtenue) = **V** (vraie valeur) + **Es** (erreur systématique) + **Ea** (erreur aléatoire)

La **vraie valeur** est la mesure « idéale », c'est-à-dire celle qui correspondrait parfaitement au phénomène étudié. Elle est le plus souvent impossible à atteindre. L'**erreur systématique** (ou biais) provient du fait que l'instrument de mesure peut présenter un écart systématique avec le phénomène étudié (par exemple, un biais lié au manque de clarté de l'échelle, à une surcharge du questionnaire, etc.). L'**erreur aléatoire** provient du fait que le phénomène mesuré par l'instrument peut être affecté par des aléas tels que la fatigue du répondant, l'humeur, etc. Ces termes d'erreur ajoutent du « bruit » aux variables observées ; la mesure obtenue contient donc à la fois la « vraie valeur » de la mesure et le « bruit ». Lorsque l'on mesurera des corrélations ou des moyennes, par exemple, l'effet mesuré sera partiellement masqué par l'erreur de mesure, ce qui entraîne un affaiblissement de l'intensité des corrélations mesurées ou une moindre précision de la moyenne calculée.

L'analyste doit donc s'interroger sur la qualité de l'instrument de mesure qu'il construit et met en œuvre. La validation d'un questionnaire, par exemple, consistera donc à tester les instruments de mesure utilisés (Hair *et al.*, 1998, p. 117-118). Ces outils de mesure doivent répondre à deux critères principaux : la **fiabilité** et la **validité**. La fiabilité renvoie à la cohérence entre les indicateurs censés mesurer le même concept, alors que la validité désigne la capacité d'un instrument de mesure à appréhender un phénomène.

- La **validité** : les instruments de mesure choisis doivent permettre d'appréhender le mieux possible le phénomène à mesurer. Il s'agit de réduire l'ensemble des termes d'erreur afin d'être en mesure de répondre à la question suivante : « Mesure-t-on bien ce que l'on cherche à mesurer ? ».
- La **fiabilité** : après s'être assuré de la validité des instruments de mesure, l'analyste peut envisager la fiabilité des mesures, en d'autres termes le fait que si l'on mesure un phénomène plusieurs fois avec le même instrument, on doit obtenir le même résultat. Il s'agit de s'assurer de la cohérence interne de l'instrument. Ce problème est concerné par l'erreur aléatoire.

De plus, après avoir diminué les erreurs de mesure par l'amélioration de chacune des variables, l'analyste doit chercher à développer des mesures multiples, autrement dit des représentations de construits cohérentes, à travers ce que l'on nomme des échelles, soit l'association de plusieurs variables dans la mesure composite d'un phénomène (voir chapitre 3).

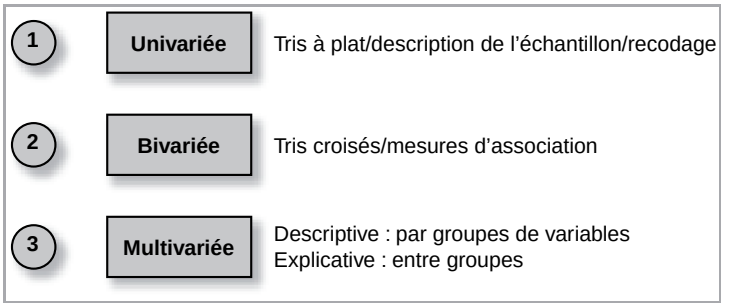
3.3 Stratégies d'analyse

L'analyse des données n'est pas une fin en soi ; elle a pour objectif d'aider à prendre une décision sur la base d'une information fiable et valide. Une stratégie d'analyse doit donc être définie afin de procéder à la modélisation d'un ou de plusieurs phénomènes. Cette stratégie repose sur la mise en lumière progressive des résultats et la complémentarité des techniques utilisées, dues à la nature des données et aux propriétés des tests envisagés. Les hypothèses qui sous-tendent les différents tests doivent être vérifiées : certaines techniques seront utiles pour étudier les différences entre variables, d'autres pour mettre en évidence leur dépendance, d'autres encore visent à classer les individus, etc. Toutes ces hypothèses seront abordées lorsque nous détaillerons l'ensemble de ces tests dans les chapitres suivants.

D'une manière générale, il est possible de représenter l'ensemble de ces techniques d'analyse en trois phases successives (voir figure 1.10).

Figure 1.10

Les stratégies d'analyse.



Source : adapté de Evrard *et al.*, 2003.

L'**analyse univariée** consiste à examiner la distribution des modalités de réponse pour une variable : dans le cas d'une variable nominale, par exemple, il s'agit d'un tri à plat, c'est-à-dire le dénombrement des observations correspondant à chaque modalité de la variable. L'**analyse bivariée** consiste à étudier les relations entre deux variables. Dans le cas de variables nominales, il s'agira d'un tableau croisé dénombrant les nombres d'observations correspondant à chaque combinaison possible des deux variables, ou plus généralement de mesures d'association quantifiant la relation (par exemple coefficient de corrélation pour des variables métriques). L'**analyse multivariée** permet de dépasser les techniques précédentes en ce sens qu'elle laisse de côté la parcellisation de l'information induite par ces techniques. En effet, si le nombre de variables est élevé, il est difficile de prendre en compte l'ensemble des combinaisons possibles. L'analyse multivariée permet donc le traitement simultané de plusieurs variables.

L'ambition du chargé d'étude quant au traitement statistique peut se situer à deux niveaux :

- **décrire les données** : il s'agira par exemple de décrire une variable (moyenne, tris), de rechercher des différences entre les modalités d'une ou de plusieurs variables (test statistique) ou encore de synthétiser et de visualiser un ensemble d'informations (analyse factorielle, typologie par exemple) ;
- **expliquer les données** : chaque méthode a ses exigences spécifiques en matière de propriété des variables (voir tableau 1.4).

Tableau 1.4 : Panorama des méthodes envisageables

Méthodes descriptives

	Variables		
	Nominale	**Ordinale**	**Métrique**
Variable	Analyse factorielle des correspondances		Analyse factorielle
Individu	Typologie		

Méthodes explicatives

	Variables indépendantes		
Une variable dépendante	**Nominale**	**Ordinale**	**Métrique**
Nominale	Logit		Régression logistique Analyse discriminante
Ordinale	Analyse conjointe	Logit ordonné	
Métrique	Analyse de variance		
Plusieurs variables dépendantes	**Nominale**	**Ordinale**	**Métrique**
Nominale	Modèles log-linéaires		
Métrique			Équations structurelles

D'une manière générale, on peut classer les méthodes selon trois dimensions :

- **descriptif/explicatif** : c'est la dimension principale en ce qui nous concerne, et celle qui structure les chapitres suivants. Les méthodes descriptives ont pour but de représenter les données ou les observations (fréquences et tris croisés : chapitre 2 ; analyses factorielles : chapitre 3 ; typologie : chapitre 4), tandis que les méthodes explicatives ont pour objet la modélisation, autrement dit la liaison entre deux phénomènes (analyse de variance : chapitre 5 ; régressions : chapitre 6 ; analyse conjointe : chapitre 7). Plus précisément – et c'est la raison pour laquelle nous faisons le lien ici avec ce qui a été évoqué lorsque nous avons précisé la notion de variable –, les méthodes explicatives traitent des relations entre deux sous-ensembles de variables : les variables à expliquer, dont on cherche à déterminer les variations ; les variables explicatives, qui contribuent à cette explication.
- **types de variables traitées** : cette dimension doit être prise en compte très en amont, lors de la création de l'instrument de mesure. En effet, le recueil de données impose automatiquement une contrainte quant aux traitements envisageables, lesquels doivent donc être anticipés. Le tableau 1.4 reprend bien les méthodes envisageables suivant les types de variables traitées. Il est important de garder à l'esprit que ces méthodes d'analyse de données ont été essentiellement développées dans des disciplines où les données sont majoritairement métriques. Le marketing reposant sur des variables principalement non métriques, il a été nécessaire d'adapter les méthodes d'analyse en introduisant des variables nominales dans des méthodes utilisant habituellement des variables métriques (régression avec variables binaires) ou en créant des méthodes utilisant ces variables qualitatives comme, par exemple, l'analyse des correspondances très populaire en marketing ;
- **nombre de variables traitées** : les méthodes descriptives ne sont limitées en termes de variables à traiter que par les capacités des logiciels utilisés. Les outils récents comme

SPSS ou SAS permettent de traiter de très larges volumes de données, et un nombre très élevé de variables. La plupart des méthodes explicatives ne peuvent traiter qu'une seule variable dépendante (ou à expliquer). Seules l'analyse canonique, l'analyse discriminante multiple et les analyses multivariées de la variance (MANOVA) et de la covariance (MANCOVA) permettent de traiter plusieurs variables dépendantes. Ces dernières seront abordées dans le chapitre 5.

Nous pouvons compléter notre tour d'horizon de la mise en œuvre des principales méthodes d'analyse de données envisageables en marketing par quelques critères complémentaires :

- l'**accent sur les individus ou les variables** (la typologie est plus fréquemment utilisée pour classer des individus alors que l'analyse factorielle est associée aux variables) ;
- la **linéarité** (la régression par exemple implique des hypothèses de linéarité des relations entre les variables étudiées, alors que la segmentation ou la typologie peuvent s'affranchir de ces hypothèses) ;
- l'aspect **paramétrique** ou **non paramétrique** (on suppose dans de nombreux tests que les variables suivent des lois de distribution théoriques dépendant d'un nombre fini de paramètres – loi normale ou multinormale par exemple ; le chargé d'étude devra donc s'assurer que ces hypothèses implicites ont été satisfaites avant de réaliser les tests).

Résumé

L'analyse de données en marketing doit être au service de la prise de décision. Ce qui implique de respecter une démarche rigoureuse faisant le lien entre le problème qui se pose au décideur et la méthode à mettre en œuvre. Cette démarche de recherche, de collecte, d'analyse et d'interprétation de l'information définit un marketing plus analytique, orienté vers une logique d'optimisation (Lilien *et al.*, 2007). Aujourd'hui, le chargé d'études dispose d'un vaste éventail de méthodes, Internet ayant fait évoluer la place traditionnellement dévolue aux approches quantitative et qualitative – approches qui semblent désormais converger grâce, notamment, à l'importance nouvelle accordée aux données secondaires. Une fois les données collectées, l'analyste doit prendre en compte les éléments liés à la précision de la mesure qu'il souhaite développer, afin de construire un instrument fiable et valide. Il mettra ensuite en place une stratégie d'analyse reposant sur la mise en lumière progressive des résultats et la complémentarité des techniques utilisées, dues à la nature des données et aux propriétés des tests envisagés.

Pour aller plus loin

Sur les études de marché : Giannelloni J. C., Vernette E., *Les Études de marché*, Broché, Paris, 2001.

Sur la démarche de recherche en marketing : Evrard Y., Pras B., Roux E., *Market. Études et recherche en marketing*, Nathan, Paris, 2003. Malhotra N., Decaudin J. M., Bouguerra A., *Études marketing avec SPSS*, 5e éd., Pearson Education, Paris, 2007.

Exercices

Exercice 1 Quand Pampers collecte des données

Énoncé

À Scwallbach, près de Francfort en Allemagne, plus de 1500 mères de famille fréquentent chaque semaine le centre d'innovation de Procter & Gamble. Elles viennent prendre des couches pour les tester et remplissent, en échange, des questionnaires. Dans l'espace de jeu à disposition, des chercheurs étudient les attitudes et comportements de bébés venus s'y amuser une partie de leur journée. Plus loin, des pièces au sol très mou – pour simuler la marche d'un tout petit –, et aux meubles géants, mettent les salariés du groupe dans la peau de jeunes enfants à différents stades de leur évolution. En France, les salariés en charge du marketing peuvent faire des « séjours d'immersion » dans des familles avec bébés, se levant la nuit avec les parents. Cette approche visant à scruter les usages et leur évolution s'inscrit dans une nouvelle démarche qui commence avec la traditionnelle boîte remise à la maternité. Des mailings prennent ensuite le relais. Les parents d'un premier enfant sont en général avides d'informations : un site internet de la marque Pampers met en avant conseils et données, des jeux en ligne – très appréciés – pour se mettre dans la peau d'un bébé, nourrissant débats, échanges, autant d'informations étudiées de près par les spécialistes de la marque.

1. Quel est le principal intérêt de la démarche de Pampers ? Quelle est la méthode utilisée, et quels en sont les principaux avantages ?
2. Comment, à votre avis, les équipes de Pampers valorisent-elles les données collectées ? Selon vous, à quels outils d'analyse ont-elles recours ?
3. Quel type de méthode, complémentaire, pourraient-elles mettre en place ? Argumentez.

Solution

1. Le principal intérêt de la démarche de Pampers est de mettre le consommateur au centre du processus de collecte de données. La méthode utilisée est à l'évidence qualitative. Elle permet d'étudier en profondeur les processus de consommation en interrogeant les parents, plus particulièrement les mères, et en observant les interactions mère-enfant. De plus, en simulant les attitudes et les comportements des bébés, elle permet surtout d'immerger les salariés du groupe dans la peau de jeunes enfants à différents stades de leur évolution. Nous sommes ici dans une démarche *orientée marché*, où la dissémination de l'information sur les consommateurs vers l'ensemble de l'organisation tient une place importante. Deux méthodes sont donc principalement utilisées : une expérimentation à Scwallbach, où les chercheurs peuvent observer et tester les comportements des bébés ; l'ethnographie en France, où les marketeurs font des séjours d'immersion dans des familles. Des outils quantitatifs d'enquête classiques prennent ensuite le relais à partir des données collectées dans les maternités.

2. Deux dimensions doivent être prises en considération. Les données issues des approches qualitatives font l'objet d'analyses de plusieurs ordres : des analyses de contenu par exemple, afin de faire émerger des thèmes, des discours, mais également un traitement des

données issues de l'expérimentation où il s'agit d'observer l'impact sur certaines variables d'une variable dont on contrôle les effets. Pour traiter des données d'expérimentation, on pourra utiliser l'analyse de variance (voir chapitre 4) ou l'analyse conjointe (voir chapitre 7), par exemple, en fonction des contraintes liées à la nature des variables.

3. Les données issues de la campagne de marketing direct et provenant du site de marque sont d'une grande richesse et peuvent nourrir de nombreuses analyses. On peut étudier les retours de la campagne de marketing direct en mettant en relation les profils sociodémographiques des parents ayant reçu la boîte d'échantillons avec la probabilité d'achat. En ce qui concerne le site internet, les *fichiers logs*, ou fichiers regroupant l'ensemble des événements survenus sur un serveur, peuvent servir de base à des analyses poussées, comme le fait Amazon.com pour customiser sa page d'accueil en fonction des profils de navigation des internautes.

EXERCICE 2 L'AUDIENCE DE LA SUPER STAR

Énoncé

Reprenons l'exemple de la mesure d'audience utilisée pour illustrer l'intervalle de confiance. Un sondeur réalise une étude d'audience par téléphone pour connaître les caractéristiques sociodémographiques et les comportements des téléspectateurs de la *Super Star*, émission de télé-réalité diffusée en prime time sur le câble et le satellite. Il sélectionne 1 000 numéros de téléphone par tirage aléatoire simple dans la base de données des abonnés de la chaîne (qui en compte 120 000 sur le câble et 2 100 000 sur le satellite). On pose l'hypothèse que les 1 000 personnes répondent effectivement aux enquêteurs. On constate que l'émission absorbe 36,8 % de l'audience des personnes interrogées de moins de 35 ans, et que le montant moyen dépensé par cette cible en SMS et appels téléphoniques est de 6,2 €, avec un écart type de 2,2 €.

1. Quel aurait été l'intervalle de confiance si l'étude d'audience avait porté sur 5 000 abonnés de la chaîne ?
2. Un annonceur souhaite investir en devenant sponsor de l'émission à condition qu'elle réalise 40 % d'audience sur les moins de 35 ans. Lui recommanderiez-vous l'investissement publicitaire ?

Solution

1. Si l'on avait interrogé 5 000 abonnés de la chaîne, on aurait calculé l'intervalle de confiance de la manière suivante :

$$p = 0{,}368$$

$$q = 1 - p = 0{,}632$$

$$0{,}368 - 1{,}96\sqrt{\frac{0{,}368 - 0{,}632)}{5000}} \leq \pi \leq 0{,}368 + 1{,}96\sqrt{\frac{(0{,}368 - 0{,}632)}{5000}}$$

$$\text{Soit : } 35{,}4\ \% \leq \pi \leq 38{,}1\ \%$$

Le sondage réalisé permet donc d'estimer cette proportion avec une précision absolue de 2,99 % (au degré de confiance 0,95). En augmentant la taille de l'échantillon, on diminue l'amplitude de l'intervalle de confiance.

2. Dans le cadre de la première étude d'audience, l'intervalle de confiance se situait entre 33,8 % et 39,7 % (au degré de confiance 0,95). La borne supérieure restant en deçà de la mesure plancher souhaitée par l'annonceur, il n'est donc pas souhaitable de réaliser l'investissement publicitaire. Il peut être intéressant de refaire le calcul pour un degré de confiance plus faible, à 0,90 (z = 1,64), à titre d'illustration. On obtient alors les résultats suivants :

$$0,368 - 1,64\sqrt{\frac{0,368-0,632}{1000}} \leq \pi \leq 0,368 + 1,64\sqrt{\frac{0,368-0,632}{1000}}$$

Soit : 34,3 % $\leq \pi \leq$ 39,3 %

Exercice 3 L'enquête « point de vente »

Énoncé

Reprenons l'exemple sur les points de vente que nous avons utilisé dans la section 2.2 (pointdevente.sav). Si l'on résume l'ensemble des questions de l'enquête dans le tableau suivant, on obtient :

- Fréquentez-vous ce point de vente au moins toutes les deux semaines ?
- Quel montant moyen dépensez-vous par mois dans ce type de point de vente ?
- Seriez-vous prêt à faire vos achats dans ce (nouveau) point de vente ?
- À combien estimez-vous le prix moyen d'une paire de chaussures dans ce point de vente ?
- Vous décririez-vous comme un auditeur régulier de radio ?
- Quel type de programme de radio écoutez-vous le plus souvent ?
- Regardez-vous régulièrement le journal télévisé ?
- Quel journal TV regardez-vous le plus fréquemment ?
- Lisez-vous la presse quotidienne ?
- Quelle rubrique de presse quotidienne lisez-vous le plus souvent ?
- Êtes-vous abonné à un titre de presse magazine ?
- La décoration de la boutique est importante à mes yeux.
- Je préfère un point de vente situé à moins de 30 minutes de chez moi.
- Je préfère être conseillé(e) par des vendeurs(ses).
- J'aime que les collections soient originales.
- J'aime qu'il y ait de nombreuses références dans les collections.
- J'aime qu'il y ait des marques connues dans les collections.
- Je préfère une décoration sobre.
- Je préfère une décoration sophistiquée.
- Je préfère une musique d'ambiance classique.
- Je préfère une musique d'ambiance rock.
- Quelle est votre année de naissance ?
- Quel est votre niveau d'étude ?

- Quel est votre statut marital ?
- En incluant les enfants de moins de 18 ans, quelle est la taille de votre foyer ?
- Quels sont approximativement les revenus de votre foyer ?
- Quel est votre sexe ?
- Possédez-vous une carte de fidélité de l'enseigne ?

1. Décrivez le type d'échelle associé à chacune des questions du tableau.
2. Donnez trois exemples de tests que vous pourriez mettre en œuvre à partir de ces variables.

Solution

1. Vous pouvez reprendre le tableau en y incluant le type de variable.

Fréquentez-vous ce point de vente au moins toutes les deux semaines ?	Nominale
Quel montant moyen dépensez-vous par mois dans ce type de point de vente ?	Numérique
Seriez-vous prêt à faire vos achats dans ce (nouveau) point de vente ?	Échelle métrique
À combien estimez-vous le prix moyen d'une paire de chaussures dans ce point de vente ?	Numérique
Vous décririez-vous comme un auditeur régulier de radio ?	Nominale
Quel type de programme de radio écoutez-vous le plus souvent ?	Nominale (échelle)
Regardez-vous régulièrement le journal télévisé ?	Nominale
Quel journal TV regardez-vous le plus fréquemment ?	Nominale (échelle)
Lisez-vous la presse quotidienne ?	Nominale
Quelle rubrique de presse quotidienne lisez-vous le plus souvent ?	Nominale (échelle)
Êtes-vous abonné à un titre de presse magazine ?	Nominale
La décoration de la boutique est importante à mes yeux.	Échelle métrique
Je préfère un point de vente à moins de 30 minutes de chez moi.	Échelle métrique
Je préfère être conseillé(e) par des vendeurs(euses).	Échelle métrique
J'aime que les collections soient originales.	Échelle métrique
J'aime qu'il y ait de nombreuses références dans les collections.	Échelle métrique
J'aime qu'il y ait des marques connues dans les collections.	Échelle métrique
Je préfère une décoration sobre.	Échelle métrique

Je préfère une décoration sophistiquée.	Échelle métrique
Je préfère une musique d'ambiance classique.	Échelle métrique
Je préfère une musique d'ambiance rock.	Échelle métrique
Quelle est votre année de naissance ?	Numérique
Quel est votre niveau d'étude ?	Nominale (échelle)
Quel est votre statut marital ?	Nominale (échelle)
En incluant les enfants de moins de 18 ans, quelle est la taille de votre foyer ?	Numérique
Quels sont approximativement les revenus de votre foyer ?	Nominale (échelle)
Quel est votre sexe ?	Nominale
Possédez-vous une carte de fidélité de l'enseigne ?	Nominale

2. De nombreux tests sont envisageables :

   a. un tri croisé entre le montant moyen dépensé dans le point de vente et le niveau d'études par exemple, afin de mettre en évidence un impact de la CSP sur les achats ;

   b. une analyse typologique afin de classer les individus de l'enquête en fonction de leur profil de réponse ;

   c. une analyse de variance multiple (MANOVA) dont l'objet serait d'expliquer le montant moyen dépensé par une série de variables explicatives comme, par exemple, le niveau d'études, le statut marital, etc.

Chapitre 2

Décrire les données

La description des données est une étape importante de la démarche d'analyse. La plupart des enquêtes se limitent à cette étape, qui donne un premier niveau de lecture des résultats ou l'identification de certaines relations entre des variables de l'étude. Cette étape peut servir de fondement, d'une part, à des analyses plus poussées, dont l'objectif est de simplifier les données (analyses factorielles par exemple), de les classer (typologies), d'autre part, à des méthodes plus sophistiquées, de nature explicative (régressions, analyses de variance, analyse conjointe, etc.). Ce chapitre a pour objectif de présenter les principales méthodes de description des données afin de produire une première analyse de ces données collectées lors d'une enquête. Après avoir abordé la nature des variables, nous étudierons les tris croisés et les principaux tests statistiques associés, ainsi que les tests d'hypothèses paramétriques et non paramétriques.

1 Description d'une variable

On appelle « variable » l'ensemble des valeurs observées sur les différents individus pour une caractéristique donnée (Tenenhaus, 1996). Une variable est qualitative dès lors qu'elle a pour valeur des modalités ; elle peut être nominale (lorsque l'ensemble des modalités ne possède pas de structure particulière) ou ordinale (lorsque l'ensemble des modalités est ordonné). Une variable est considérée comme quantitative ou métrique lorsque ses modalités peuvent être mesurées (par exemple l'âge, la valeur d'une action, etc.).

1.1 Décrire une variable qualitative

La description d'une variable qualitative consiste à présenter les **effectifs**, c'est-à-dire le nombre d'individus de l'échantillon pour chaque modalité de la variable, et les **fréquences**, c'est-à-dire le nombre de réponses associées aux modalités de la variable étudiée. En effet, dans de nombreux cas, le chargé d'étude cherche à répondre à une série de questions ne concernant qu'une seule et même variable.

SPSS

Il existe plusieurs possibilités dans SPSS pour décrire les données collectées. On peut par exemple, dans un premier temps, générer un rapport sur les observations pour s'assurer qu'elles ne comportent pas d'erreurs de saisie, de valeurs aberrantes (**Analyse > Rapport > Récapitulatif des observations…**) ou plus simplement pour prendre connaissance des variables dans un tableau synthétique, ce qui s'avère souvent utile en début d'analyse (**Outils > variables…**).

La procédure *Fréquence* permet d'obtenir les affichages statistiques et graphiques qui servent à décrire des variables quantitatives et qualitatives. Pour obtenir un tableau d'effectifs et de fréquences pour une ou plusieurs variables dans SPSS, ouvrez le fichier de données « pointdevente.sav », sélectionnez dans le menu **Analyse > Statistiques descriptives > Effectifs…**, puis procédez à la description de la variable de type nominal *marital* correspondant à la question : « Quel est votre statut marital ? ». La boîte de dialogue de la figure 2.1 apparaît.

Figure 2.1

Boîte de dialogue de la procédure *Fréquence*.

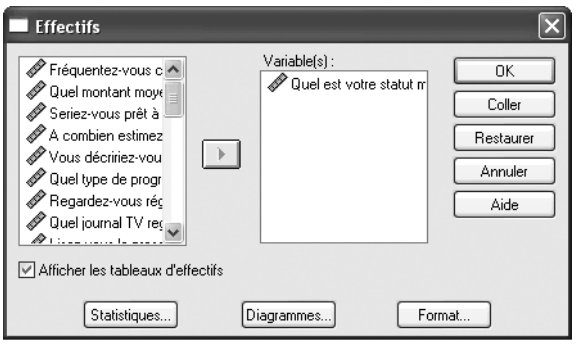


Figure 2.2

Description de la variable *marital*.

Quel est votre statut marital?

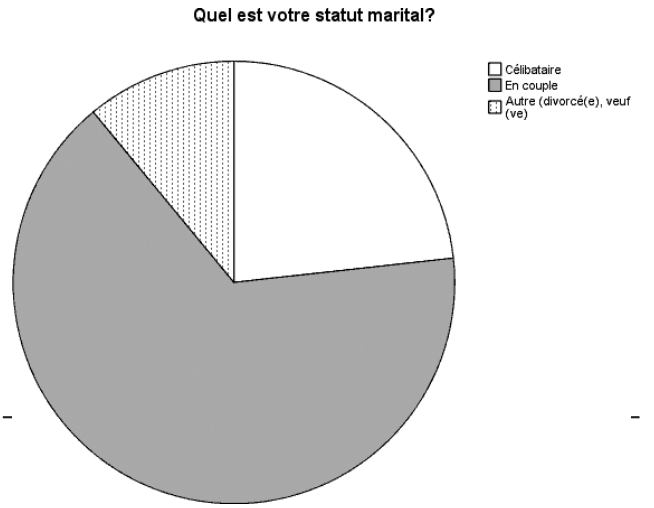
		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Célibataire	93	23,3	23,3	23,3
	En couple	263	65,8	65,8	89,0
	Autre (divorcé(e), veuf(ve)	44	11,0	11,0	100,0
	Total	400	100,0	100,0	

La figure 2.2 correspond à un **tri à plat** de la variable qualitative *marital ;* en d'autres termes, il reprend les effectifs et les fréquences (présentés ici en pourcentage) pour une variable. L'intérêt du tri à plat est de fournir une description rapide de la variable étudiée. Le tableau montre immédiatement que 65,8 % des individus de l'échantillon interrogé sont en couple et que 23,3 % sont célibataires.

Ces résultats peuvent également être visualisés sous forme de graphiques (diagrammes en bâtons, en secteurs), dans lesquels les surfaces associées aux différentes modalités sont proportionnelles à leur fréquence, exprimée en valeur ou en pourcentage, comme le montre la figure 2.3.

Figure 2.3

Diagramme en secteurs des effectifs de la variable *marital*.



1.2 Décrire une variable quantitative

Plusieurs critères permettent de décrire une variable quantitative :

- **les mesures de la tendance centrale** : moyenne, médiane, mode ;
- **les mesures de la dispersion** : étendue, variance, écart type, coefficient de variation ;
- **les mesures de la distribution** : asymétrie, aplatissement ;
- **les représentations graphiques** : histogrammes ou boîtes à moustaches, par exemple.

Mesures de la tendance centrale

Les **mesures de la tendance centrale** ont pour objet de résumer la série d'observations par une valeur considérée comme *représentative*. La plus fréquemment employée est la **moyenne**, ou somme des valeurs de toutes les observations divisée par l'effectif ; celle que l'on utilise le plus souvent est la moyenne arithmétique. La moyenne révèle la tendance centrale en ce sens que les réponses se trouvent réparties de part et d'autre de la moyenne. Si certaines valeurs sont très éloignées les unes des autres, elles peuvent avoir une influence importante sur la moyenne. Dans ce cas, il vaut mieux utiliser la médiane, qui n'est pas sensible aux valeurs aberrantes ou extrêmes (*outliers*). La **médiane** représente la valeur au-dessus et au-dessous de laquelle se situent la moitié des observations, c'est-à-dire le 50e centile (voir focus 2.1 ci-après). Le **mode** représente la valeur présentant la plus grande fréquence d'occurrence. Si plusieurs valeurs à la fois présentent la plus grande fréquence d'occurrence, chacune d'entre elles est un mode.

Focus 2.1 Les fractiles

Les fractiles sont les valeurs d'une variable quantitative qui divisent les données triées en classes par centième. Les quartiles (25e, 50e et 75e centiles) divisent les observations en quatre classes de taille égale. On les définit dans SPSS à partir de la boîte de dialogue **Effectifs > Statistiques** (voir figure 1.1), en sélectionnant **Partition en** *n classes égales* (*n* définissant le niveau de partition souhaité). Vous pouvez également spécifier des centiles particuliers (par exemple le 95e centile), autrement dit les valeurs au-dessus de 95 % des observations.

Mesures de la dispersion

Les **mesures de la dispersion** reposent sur les indicateurs suivants : l'étendue, la variance, l'écart type et le coefficient de variation. L'**étendue** (ou **intervalle**) est la différence entre la plus grande et la plus petite des valeurs observées. La **variance** est la mesure de la dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un. Lorsque les données se concentrent autour de la moyenne, la variance est faible. Si les données sont dispersées autour de la moyenne, la variance est élevée. Il s'agit d'une mesure plus fine de la dispersion, au sens où toutes les données sont prises en compte. En revanche, elle est sensible aux valeurs extrêmes. L'**écart type** est la mesure de la dispersion autour de la moyenne, exprimée dans la même unité que la variable. L'écart type est la racine carrée de la variance. On l'écrit de la manière suivante :

$$s = \sqrt{\frac{\sum_{1=1}^{n} (Xi - \bar{X})\dagger}{n-1}}$$

Le **coefficient de variation** est le rapport de l'écart type à la moyenne ($CV = \frac{s}{\bar{X}}$), exprimé en pourcentage. Son objet est de mesurer le degré de variation de la moyenne d'un échantillon à l'autre, lorsque ceux-ci sont issus de la même distribution.

Mesures de la distribution

On **mesure la symétrie et la forme de la distribution** par l'asymétrie et l'aplatissement. Ces statistiques sont présentées avec leur erreur standard.

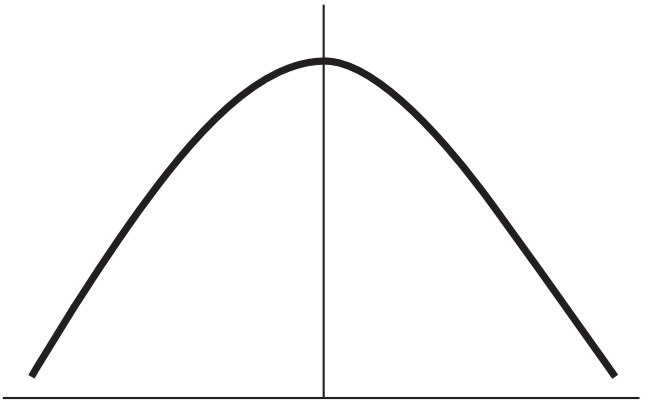
Le **coefficient de symétrie (skewness)** mesure l'asymétrie d'une distribution. Une distribution normale est symétrique (voir figure 2.4), c'est-à-dire que les valeurs sont les mêmes de part et d'autre du centre de la distribution, et possède une valeur de skewness de 0. Une distribution avec un skewness positif significatif est une distribution asymétrique à droite (la distribution prend la forme d'une longue queue à droite) et une distribution avec un skewness négatif significatif est une distribution asymétrique à gauche (la distribution prend la forme d'une longue queue à gauche). Cette asymétrie s'explique par le fait que les écarts sont plus importants dans une direction que dans l'autre.

Le **coefficient d'aplatissement (kurtosis)** permet de mesurer le relief ou la platitude d'une courbe issue d'une distribution de fréquences. En d'autres termes, le coefficient d'aplatissement permet de mesurer le degré de concentration des observations dans les queues de la courbe. Le coefficient de kurtosis est de 0 pour une distribution normale (gaussienne).

Un kurtosis négatif indique donc que les queues comptent un plus grand nombre d'observations que dans une distribution gaussienne. Les coefficients de kurtosis et de skewness peuvent être utilisés pour s'assurer que les variables suivent une distribution normale, condition nécessaire pour de nombreux tests statistiques. On estime que le coefficient de symétrie ou skewness doit être inférieur à 1 et le coefficient d'aplatissement ou kurtosis doit être inférieur à 1,5 pour considérer que la variable suit bien une loi normale.

Figure 2.4

Représentation d'une distribution normale.



$$Mo = Me = \overline{X}$$

SPSS

Reprenons notre exemple avec SPSS (pointsdevente.sav) : rappelez la boîte de dialogue de la procédure précédente (**Effectifs**) en cliquant sur l'icône dans la barre d'outils. Procédez aux mêmes opérations mais cette fois pour la variable *montant*. Dans la boîte de dialogue **Effectifs** que vous venez de rappeler, cliquez sur l'onglet **Statistiques** et cochez les statistiques de mesure de la tendance centrale, de dispersion et de distribution, puis sélectionnez un graphique (un histogramme avec courbe gaussienne par exemple) pour représenter la distribution.

Les figures 2.5 et 2.6 reprennent les statistiques descriptives de la variable *montant*.

Figure 2.5

Description de la variable *montant*.

Statistiques

Quel montant moyen dépensez-vous par mois dans ce type de point de vente?

N	Valide	400
	Manquante	0
Moyenne		153.5100
Erreur std. de la moyenne		4.55739
Médiane		172.0000
Mode		.00
Ecart-type		91.14782
Variance		8307,925
Asymétrie		-,067
Erreur std. d'asymétrie		,122
Aplatissement		-,085
Erreur std. d'aplatissement		,243
Intervalle		444.00
Minimum		.00
Maximum		444.00
Somme		61,404.00

Figure 2.6

Représentation d'un graphique de la variable *montant*.

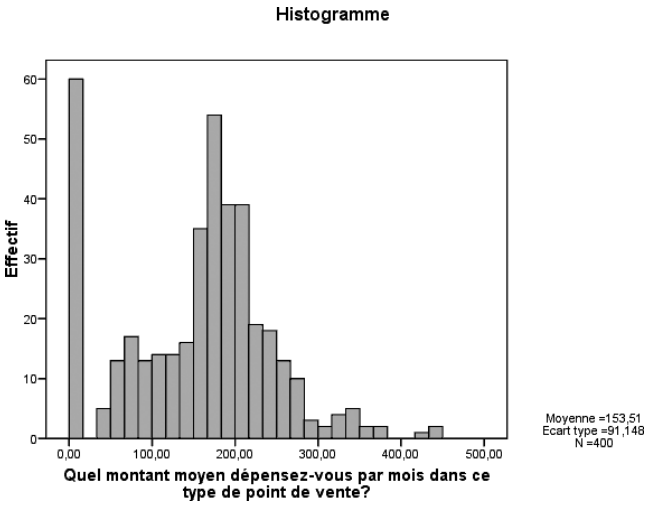
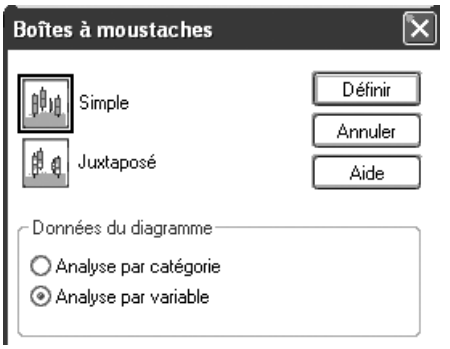


Figure 2.7

Création d'une boîte à moustaches.



Le montant moyen dépensé dans le point de vente est de 153,51 €, avec un écart type de 91,15 €. Pour 59 répondants, le montant est nul, c'est-à-dire qu'il s'agit de non-clients du magasin. En termes de dispersion, la variance est élevée (8 307,9) en raison de valeurs extrêmes importantes, ce qui est confirmé par l'écart type. On constate que l'asymétrie pour la variable *montant* est légèrement négative (–0,67).

Représentations graphiques

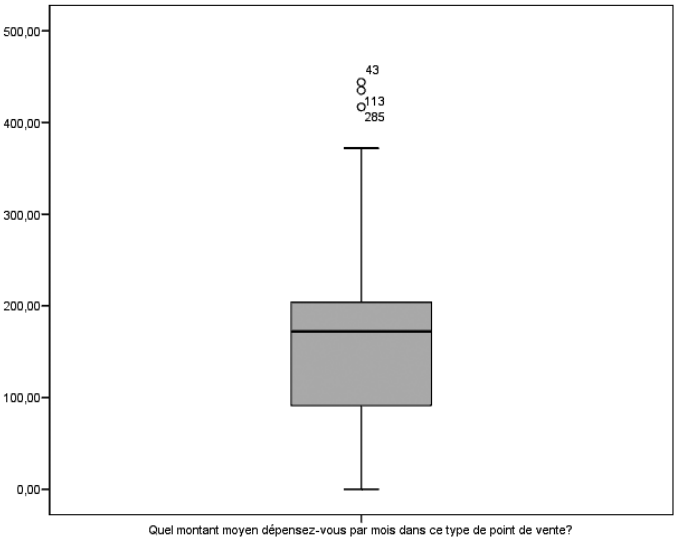
En ce qui concerne les **représentations graphiques**, les fréquences peuvent être représentées par des histogrammes et des graphiques en secteurs, comme nous l'avons vu précédemment. Pour visualiser la répartition des fréquences, les diagrammes en bâtons sont souvent pertinents.

La réalisation des graphiques dans SPSS s'effectue soit à partir des boîtes de dialogue des différents tests (dans notre cas, le menu **Effectifs**), soit directement dans le menu **Graphes**. Parmi les options qui vous sont proposées, sélectionnez **Boîtes de dialogues héritées** dans le menu **Graphes**, puis de nouveau la variables *montant*. Sélectionnez le graphique **Boîte à moustaches**, puis, dans **Données du diagramme**, l'option **Analyse par variable** (voir figure 2.7).

La boîte à moustaches est une représentation graphique intéressante car elle permet de récapituler une variable numérique en représentant la médiane, les quartiles et les valeurs extrêmes. Cliquez sur **Définir** : on vous propose d'étiqueter les observations en utilisant une variable de type numérique ou une variable textuelle afin d'identifier les valeurs extrêmes. Si vous ne choisissez rien, les numéros d'observation serviront à étiqueter ces valeurs. Nous obtenons le graphique représenté à la figure 2.8.

Figure 2.8

Représentation de la variable *montant* sous forme de boîte à moustaches.



L'intérêt de cette représentation est qu'elle permet de visualiser de manière compacte la dispersion des données. La figure 2.8 montre des valeurs extrêmes qui apparaissent isolées du graphique. On peut donc observer que le montant dépensé varie entre 444 € (observation n° 43) et 0 € (moustache inférieure), avec une médiane qui partage la boîte centrale et qui est de 172 €.

Il est possible d'aller plus loin dans la description des variables en sélectionnant les observations sur lesquelles on souhaite faire porter l'analyse. On peut notamment chercher à savoir si les hommes dépensent en moyenne plus ou moins que l'ensemble de la population. Pour ce faire, il faudra filtrer les observations en fonction du sexe des répondants. Dans le menu **Données**, appelez la boîte de dialogue **Sélectionner les observations** puis, dans la partie **Sélectionner**, cliquez sur **Selon une condition logique**. Pour ne sélectionner que les hommes, vous devez faire glisser la variable *sexe* en précisant la condition : « sexe = 1 » (1 étant l'étiquette retenue pour les hommes). Vous obtenez la boîte de dialogue de la figure 2.9.

Figure 2.9

Boîte de dialogue Sélectionner des observations.

Dans la fenêtre de résultats, on obtient un montant moyen dépensé par les hommes de 155,89 €, avec un écart type de 95,31 €, montants légèrement supérieurs à la dépense moyenne de l'échantillon. On remarque également que les hommes représentent un peu plus de la moitié des répondants (204 observations).

2 Analyses bivariées

L'examen de variables uniques permet une première lecture intéressante des résultats mais elle ne présente pas de véritable intérêt en termes d'analyse. Les descriptions faites sur les variables soulèvent toute une série de questions sur leurs relations, qui devront être mises en lumière en les rapprochant deux à deux dans des **analyses bivariées**. Les **tris croisés**, par exemple, permettent d'examiner les relations entre deux ou plusieurs variables. Ces relations peuvent être *symétriques* – l'analyse cherche à mesurer la liaison entre les deux variables et à en tester la signification –, ou *dissymétriques* – l'analyse cherche à expliquer les variations d'une variable dépendante par les variations d'une variable indépendante (Evrard *et al.*, 2003). Ce dernier cas constituant le plus souvent une occurrence particulière des méthodes multivariées explicatives (corrélations, ANOVA, etc.), il sera traité dans les chapitres suivants.

2.1 Tris croisés

Les tableaux croisés à deux ou plusieurs modalités sont en général complétés par des mesures d'association qui permettent de démontrer la signification statistique d'une association observée entre les variables. Ces tests seront développés dans la section suivante.

Les **tris croisés** ont pour objet de rassembler dans un tableau unique les distributions de fréquences de deux ou plusieurs variables. Ce premier outil d'analyse des relations entre deux variables, ou relations bivariées, permet de répondre à des questions qui se posent dès l'origine de l'étude (par exemple : « Les hommes dépensent-ils plus que les femmes sur le point de vente ? » ; « Le sexe et les revenus ont-ils une influence sur le montant moyen dépensé ? ») ou de mettre en lumière des relations dont on soupçonne l'existence à l'issue des traitements réalisés variable par variable. Le principe du tableau croisé est de proposer une ventilation des fréquences de réponse par variable et par modalité (voir figure 2.10).

SPSS

Il existe deux approches pour générer un tableau croisé dans SPSS. Vous pouvez créer un tableau croisé depuis le menu **Analyse > Statistiques descriptives > Tableaux croisés…** ou bien depuis le menu **Analyse > Tableaux > Tableaux personnalisés…** Nous utiliserons ici la seconde possibilité. Pour ventiler les montants moyens dépensés en fonction du sexe – nous avons déjà obtenu les données variable par variable –, faites glisser la variable *montant* de la liste des variables vers la zone Lignes du tableau. L'unité d'analyse proposée par défaut est la moyenne, la variable étant métrique. Puis faites glisser la variable *sexe* de la liste vers la zone Colonnes du tableau.

Figure 2.10

Tri croisé du montant moyen dépensé en fonction du sexe.

	Quel est votre sexe?	
	Homme	Femme
	Moyenne	Moyenne
Quel montant moyen dépensez-vous par mois dans ce type de point de vente?	155.89	151.03

Poursuivons l'exploration en introduisant une troisième variable : les revenus. L'introduction d'une troisième variable est pertinente si elle permet d'affiner l'association entre les deux variables. Rappelez la boîte de dialogue **Tableaux personnalisés** et faites glisser la variable *revenus* de la liste vers la zone Colonnes du tableau. Le tableau obtenu est relativement difficile à lire, car trop large. Double-cliquez sur le tableau obtenu dans votre feuille de résultats SPSS pour ouvrir un tableau pivotant. Le tableau pivotant vous permet d'inverser lignes et colonnes. On obtient la figure 2.11 ci-après.

Figure 2.11

Tri croisé du montant moyen dépensé en fonction du sexe et des revenus.

					Quel montant moyen dépensez-vous par mois dans ce type de point de vente?
Quels sont approximativement les revenus de votre foyer?	<15 000€	Quel est votre sexe?	Homme	Moyenne	.00
			Femme	Moyenne	.00
	15 000€ - 24 999€	Quel est votre sexe?	Homme	Moyenne	.14
			Femme	Moyenne	.00
	25 000€ - 49 999€	Quel est votre sexe?	Homme	Moyenne	124.08
			Femme	Moyenne	122.31
	50 000€ - 74 999€	Quel est votre sexe?	Homme	Moyenne	133.02
			Femme	Moyenne	136.00
	75 000€ - 99 999€	Quel est votre sexe?	Homme	Moyenne	194.97
			Femme	Moyenne	199.00
	100 000 - 149 999€	Quel est votre sexe?	Homme	Moyenne	222.04
			Femme	Moyenne	223.25
	+ de 150 000€	Quel est votre sexe?	Homme	Moyenne	259.95
			Femme	Moyenne	233.10

On constate que les montants moyens dépensés augmentent *a priori* en fonction des revenus, relation qui apparaît relativement moins évidente en fonction du sexe. Les tris croisés ne nous permettent pas de déduire quoi que ce soit quant au type de relation existant entre les variables. Avant de conclure à une éventuelle relation entre le montant moyen dépensé et les revenus ou le sexe, le chargé d'étude doit donc mesurer la force d'association entre ces variables. S'il souhaite étudier l'influence d'une variable sur une autre, il devra mettre en œuvre le test approprié (voir la section 3 du chapitre).

2.2 Tests d'association de deux variables

Les tris croisés présentent la distribution des fréquences de réponse pour deux ou plusieurs variables mises en relation mais ils ne permettent pas de démontrer l'existence de cette association du point de vue statistique. Pour mesurer véritablement la relation entre les variables, il est nécessaire de mettre en place des tests de signification statistique de l'association. Nous aborderons de manière plus précise la théorie des tests statistiques dans la section 3 de ce chapitre.

Test du khi-deux

Le test le plus couramment utilisé est celui du **khi-deux** (χ^2), car il consiste à tester la signification statistique d'une association de deux variables qualitatives (nominales ou ordinales). Plus précisément, il a pour objet de tester l'indépendance des variables dans un tableau croisé en comparant la distribution observée (O_{ij}) sur l'échantillon à une distribution théorique (T_{ij}) qui correspond à l'hypothèse que l'on veut tester. Le χ^2 observé sur l'échantillon se calcule de la manière suivante :

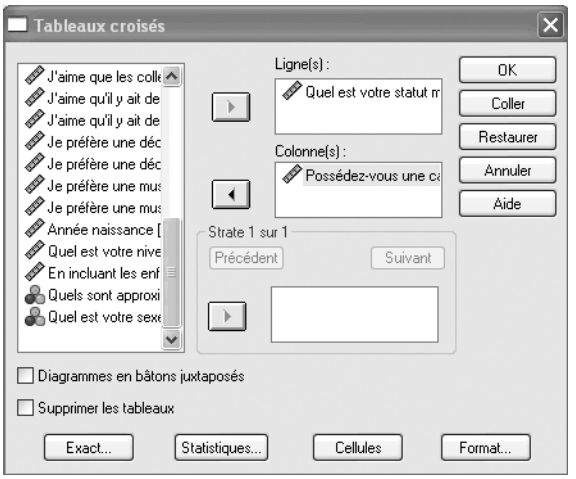
$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - T_{ij}\right)}{T_{ij}}$$

La loi du khi-deux suit une distribution asymétrique dont la forme dépend du nombre de degrés de liberté (DDL). Le nombre de degrés de liberté varie en fonction du nombre de modalités des variables comparées et se calcule de la manière suivante : $n - 1 \times p - 1$ (avec n : modalités de la 1re variable et p : modalités de la 2e variable). On rejettera l'hypothèse nulle (pas d'association entre les variables) si le χ^2 calculé est supérieur à la valeur de référence du χ^2 se trouvant dans la table de khi-deux pour n degrés de liberté (lignes) et pour un α (niveau de précision donné en colonnes). Pour interpréter la valeur du χ^2, il est préférable de se référer au seuil de signification statistique (> 0,05 par exemple) plutôt qu'à la valeur du χ^2 qui varie selon le nombre de degrés de liberté.

Le test du khi-deux s'obtient par la procédure des tableaux croisés vue plus haut (**Analyse** > **Statistiques descriptives** > **Tableaux croisés…**) et peut être sélectionné dans le menu **Statistiques**, comme l'indique la figure 2.12.

Figure 2.12

Boîte de dialogue du tableau croisé et test du khi-deux.



Si l'on cherche à établir le profil des clients les plus fidèles en croisant le statut marital et la possession d'une carte de fidélité, par exemple, le test du khi-deux permettra de définir si ces deux variables sont indépendantes. Il est important de noter que ce test est assez sensible à la taille de l'échantillon et que chaque case du tableau doit comporter au moins cinq observations (voir figures 2.12 et 2.13).

Figure 2.13

Tableau croisé des variables *marital/carte*.

Effectif

		Possédez-vous une carte de fidélité de l'enseigne?		
		Non	Oui	Total
Quel est votre statut marital?	Célibataire	78	15	93
	En couple	214	49	263
	Autre (divorcé(e), veuf(ve)	29	15	44
Total		321	79	400

Nous avons créé un tableau croisé dans SPSS selon la procédure présentée plus haut et sélectionné le test du khi-deux dans le menu **Statistiques** de la boîte de dialogue **Tableaux croisés**. Conformément à ce que nous pouvions penser *a priori*, la valeur du χ^2 est à la fois élevée et supérieure à la valeur critique correspondant au seuil de signification statistique de 0,05 (nous obtenons 0,035). Ce résultat nous permet de rejeter *Ho* (« il n'existe pas de lien entre les variables ») et de conclure qu'il existe bien une relation entre le statut marital et la possession d'une carte de fidélité dans la population observée.

Figure 2.14

Test du khi-deux des variables *marital/carte*.

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	6,687[a]	2	,035
Rapport de vraisemblance	5,977	2	,050
Association linéaire par linéaire	4,499	1	,034
Nombre d'observations valides	400		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 8,69.

Autres tests

Dans le cas particulier des tableaux carrés ou 2 × 2 (2 lignes et 2 colonnes), qui comparent deux variables à deux modalités, il est recommandé d'appliquer une correction au χ^2, ou d'utiliser le **coefficient phi** (ϕ). Celui-ci correspond à la racine carrée du χ^2 divisé par la taille de l'échantillon, soit :

$$\phi = \sqrt{(\chi^2/n)}$$

Le **coefficient de contingence** (C) peut être appliqué pour des mesures d'association sans contrainte de taille de tableau. L'indicateur oscille entre une borne inférieure de 0 lorsqu'il n'y a aucune association (lorsque $\chi^2 = 0$) et une borne supérieure inférieure à 1. Cette valeur maximale du coefficient dépend de la taille du tableau (nombre de lignes × nombre de colonnes), raison pour laquelle il ne doit être employé que pour comparer des tableaux de même taille. On le calcule de la manière suivante :

$$C = \sqrt{(\chi^2/\chi^2+n)}$$

Le **V de Cramer** est un coefficient normé, c'est-à-dire qu'il peut atteindre 1, quelle que soit la taille du tableau. Il s'agit d'une version modifiée du coefficient phi (ϕ). Il est noté :

$$V = \sqrt{\frac{\chi^2/n}{\min(r-c),(c-1)}}$$

Le **coefficient d'association prédictive** (**lambda**) permet de mesurer le pourcentage d'amélioration de la valeur d'une variable nominale dépendante en fonction de la valeur de la variable nominale indépendante, celle-ci étant utilisée comme pivot. Le **lambda** est un coefficient dissymétrique, c'est-à-dire que le résultat varie selon la variable qui sert de pivot (ainsi dans la régression, par exemple).

3 Théorie des tests statistiques

Les tests statistiques reposent sur le principe d'inférence, c'est-à-dire le fait de procéder à des généralisations sur les comportements d'une population. Ils sont fondés sur des mesures effectuées sur des variables ou sur des facteurs à partir d'observations réalisées sur un échantillon de cette population. L'objectif de la statistique dans la logique inférentielle est donc de tester des hypothèses formulées essentiellement sur la base d'une théorie préexistante ou de résultats antérieurs.

3.1 L'HYPOTHÈSE STATISTIQUE

Une hypothèse statistique est un énoncé quantitatif concernant les caractéristiques d'une population ou, plus précisément, une affirmation portant sur une ou plusieurs variables. Elle se présente traditionnellement sous la double forme d'une première hypothèse, appelée **hypothèse nulle**, et d'une seconde hypothèse, appelée **hypothèse alternative**. Son objectif est de réfuter l'hypothèse nulle, laquelle concerne le plus souvent un *statu quo* ou une absence de différence, au profit de l'hypothèse alternative.

Exemple : on peut poser l'hypothèse nulle Ho *qu'il n'existe pas de différence de ventes entre les points de vente situés en centre-ville et ceux de la périphérie urbaine, et l'hypothèse alternative* H1 *qu'elles sont différentes en centre-ville et en périphérie urbaine.*

Les tests statistiques étant conçus pour la réfutation d'hypothèses et non pour leur confirmation, l'hypothèse alternative est celle qui sera acceptée si l'hypothèse nulle est rejetée. Accepter une hypothèse revient donc à dire que l'hypothèse est non rejetée plutôt qu'acceptée, c'est-à-dire que les données recueillies au cours d'une expérience particulière sont compatibles avec l'hypothèse alternative proposée.

L'objectif de l'analyse de données est donc de prendre une décision : en l'occurrence, rejeter ou non l'hypothèse nulle *Ho*. Les tests étant fondés sur des informations incomplètes issues d'observations portant sur un échantillon de la population, il est nécessaire de définir le **seuil de signification** du test, seuil formulé en pourcentage de chances de rejeter l'hypothèse nulle alors qu'en réalité celle-ci était vraie. Le seuil de signification est habituellement noté α et exprimé en pourcentage. Le choix du seuil est lié au niveau de risque accepté (1 % ou 5 % étant les valeurs usuelles). Son complément $(1 - \alpha)$, appelé **seuil de confiance**, correspond au pourcentage de cas où on acceptera l'hypothèse nulle à juste titre. On appelle **erreur de type I** le fait de rejeter, à la suite des résultats d'un test statistique, une hypothèse qui serait en réalité vraie (condamner un innocent) et **erreur de type II** l'erreur liée au fait d'accepter une hypothèse qui serait en réalité fausse (innocenter un coupable). La probabilité de commettre ce type d'erreur est notée β ; on appelle **puissance du test** son complément $(1 - \beta)$, lequel correspond à la probabilité de rejeter une hypothèse qui serait réellement fausse (voir tableau 2.1).

Tableau 2.1 : Types d'erreurs dans un test statistique

		Situation dans la population	
		Ho vraie	**Ho fausse**
Décision	**Ho acceptée**	Décision correcte (seuil de confiance = $1 - \alpha$)	Erreur de type II ($\boldsymbol{\beta}$)
	Ho rejetée	Erreur de type I (seuil de signification = $\boldsymbol{\alpha}$)	Décision correcte (puissance du test = $1 - \boldsymbol{\beta}$)

Bien que l'α établisse le niveau de signification du test, c'est la puissance du test $(1 - \beta)$ qui donne une estimation de la probabilité de trouver des différences significatives – si elles existent – dans les données. Pourquoi, dès lors, ne pas prendre en compte l'α et le β en tant que niveaux de confiance ? La raison évoquée est que l'erreur de type I et l'erreur de

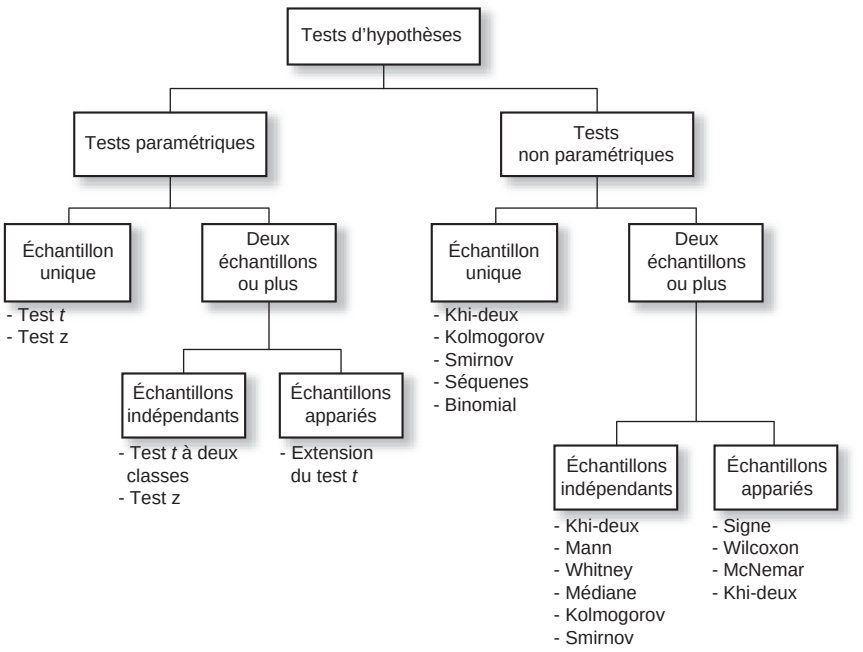
type II sont inverses : plus l'erreur de type I devient restrictive (proche de 0) et plus la probabilité d'une erreur de type II augmente ; de même, réduire l'erreur de type I réduit la puissance du test. L'analyste doit donc trouver le juste équilibre entre le degré de confiance (α) et la puissance du test qui en résulte. La seule manière de faire baisser simultanément α et β est d'augmenter la taille de l'échantillon étudié.

3.2 Les tests d'hypothèses

Les tests d'hypothèses, ou tests d'inférence, ont pour objectif de mesurer l'effet d'une variable indépendante sur une variable dépendante, en fonction du nombre d'échantillons et en fonction de la nature des variables étudiées. On nomme **tests paramétriques** les approches reposant sur des données métriques (et par suite sur des paramètres connus tels que la moyenne ou l'écart type, par exemple), et **tests non paramétriques** les approches reposant sur des données non métriques (et qui, par suite, peuvent s'affranchir de conditions de distribution particulières). Les tests non paramétriques étant peu sensibles à la taille de l'échantillon et aux données aberrantes, ils sont utilisés en marketing où les échantillons peuvent parfois être de petite taille (moins de 30 individus). Le nombre d'échantillons joue également un rôle important dans le choix du test approprié. En effet, deux situations doivent être distinguées : lorsque l'on étudie deux populations distinctes sur une même variable, on parle de **mesures indépendantes** (comparer les clients et les non-clients) ; et lorsque les mêmes individus sont mesurés sur une même variable dans deux situations distinctes, on parle de **mesures appariées** (comparer les niveaux de prix à deux périodes distinctes). Ces éléments affectent de manière importante les statistiques de tests (voir figure 2.15).

Figure 2.15

Tests paramétriques et tests non paramétriques (Malhotra *et al.*, 2007).



3.3 Tests paramétriques

Les deux principaux tests paramétriques sont le test *t* et le test *Z*, qui ont pour objet de tester des différences de moyenne. Ces tests sont souvent mis en œuvre en marketing, car ils permettent, par exemple, de comparer la moyenne d'une variable dépendante métrique en fonction des modalités d'une variable nominale. On formule alors une hypothèse nulle qui sera vérifiée par le test *t* ou le test *Z*. Pour plus de simplicité, ces deux tests sont présentés ici pour des échantillons uniques.

Test *t*

Le **test *t*** est directement lié à la statistique *t* de Student, qui suppose que la variable adopte une distribution normale, que la moyenne soit connue et que la variance, lorsqu'elle est inconnue, soit estimée sur l'échantillon. On le calcule de la manière suivante :

$$t = \left(\bar{X} - \mu\right) / s_{\bar{X}}$$

Où : $\bar{X}$: moyenne de l'échantillon
μ : moyenne de la variable
$s_{\bar{X}}$: variance de l'échantillon

Dans SPSS, ce test paramétrique peut être estimé avec la procédure suivante : menu **Analyse** > **Comparer les moyennes** > **Test T pour échantillon unique…**, procédure que nous avons utilisée au chapitre 1 pour estimer l'intervalle de confiance. Pour comparer les moyennes de deux échantillons indépendants (comparaison des clients et des non-clients par exemple), on utilisera une analyse de variance (ANOVA) à 1 facteur (voir chapitre 4). Pour comparer les moyennes de deux échantillons appariés (comparaison de relevés de prix à deux périodes distinctes par exemple), on suivra une extension du test *t* pour échantillons appariés qui est disponible dans la même boîte de dialogue.

Test *Z*

Le test *Z* peut être mis en place lorsque la variance de l'échantillon est connue. La valeur de *Z* s'obtient par la formule suivante :

$Z = \left(\bar{X} - \mu\right) / \sigma_{\bar{X}}$ où : $\sigma_{\bar{X}}$ est l'écart type de la population

Ce test peut également être étendu pour tester des proportions.

3.4 Tests non paramétriques

Les tests non paramétriques sont souvent mis en œuvre dans la pratique en marketing : ils s'appliquent aux variables qualitatives et s'avèrent relativement performants sur de petits échantillons, même s'ils sont moins puissants que les tests paramétriques. Voici les principaux tests paramétriques présentés ici : un **test d'ajustement** (le test de Kolmogorov-Smirov), des **tests de comparaison d'échantillons indépendants** (le test U de Mann-Whitney et le test de la médiane), ainsi que des **tests de comparaison d'échantillons appariés** (le test de Wilcoxon, le test du signe et le test de McNemar).

- **Test de Kolmogorov-Smirov (K-S)**

Le test de Kolmogorov-Smirov est un test dit d'ajustement, car il permet d'établir si une population donnée suit une distribution particulière (normale, uniforme ou poisson par exemple), condition exigée par de nombreux tests. Le K-S est calculé à partir de la plus grande différence (en valeur absolue) entre les fonctions de distribution théorique et observée cumulées :

$$K = \text{Max } |A_i - O_i|$$

Le K-S pour un échantillon s'obtient dans SPSS à partir du menu **Analyse > Tests non paramétriques > K-S à 1 échantillon…**

- **Test *U* de Mann-Whitney**

  Le test de Mann-Whitney permet de vérifier que deux échantillons (ou groupes) proviennent bien de la même population. On peut l'utiliser, par exemple, pour comparer les réponses dans un département par rapport aux réponses nationales. La statistique du test *U* réunit les deux échantillons et ordonne les observations par ordre croissant de taille. Le test calcule le nombre de fois où un résultat du groupe 1 précède un résultat du groupe 2, ainsi que le nombre de fois où un résultat du groupe 2 précède un résultat du groupe 1. *U* est d'autant plus petit que les groupes sont différents.

  Pour calculer le *U* de Mann-Whitney dans SPSS, il faut d'abord définir la variable qui servira à scinder les données en deux échantillons : **Analyse > Test non paramétrique > 2 échantillons indépendants…**, puis sélectionner une variable de regroupement (**Facteur**) et cliquer sur **Définir les niveaux.** Pour définir les groupes, vous devez indiquer les valeurs pour le groupe 1 et celles pour le groupe 2. Sélectionnez ensuite le **test *U* de Mann-Whitney** dans la boîte de dialogue.

- **Test de la médiane**

  Ce test, moins puissant que le *U* de Mann-Whitney, permet de déterminer si deux groupes sont issus de populations ayant la même médiane, en estimant la position de chaque observation par rapport à la médiane globale des deux échantillons.

  Pour calculer le test de la médiane dans SPSS, vous devez suivre la procédure suivante : **Analyse > Test non paramétrique > K échantillons indépendants…**, puis sélectionner le **test de la médiane** dans le menu du type de test envisagé.

- **Test de Wilcoxon**

  Le test de Wilcoxon est utilisé dans le cas de la comparaison de deux échantillons appariés, c'est-à-dire lorsque l'on souhaite, par exemple, comparer deux types de réponses : avant/après l'exposition à un message publicitaire, attitude par rapport à une marque A et une marque B, etc. La statistique *z* du test de Wilcoxon s'obtient en calculant la différence entre les scores des deux observations par paires d'observations, puis en calculant le rang de toutes les différences, et enfin la somme des rangs positifs et des rangs négatifs. On rejette l'hypothèse nulle (absence de différence entre les deux groupes) s'il y a une différence entre la somme des rangs positifs et la somme des rangs négatifs. Le sens de la statistique indique le sens de la différence de la paire examinée.

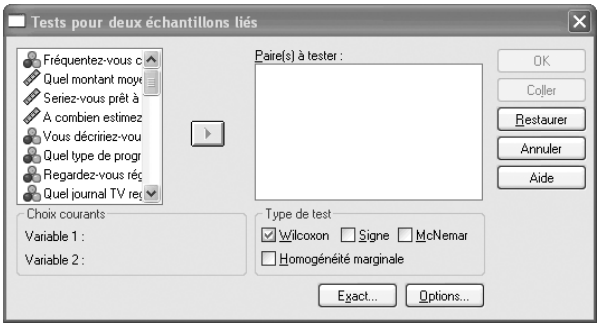
  Dans SPSS, ouvrez le menu **Analyse > Test non paramétrique > 2 échantillons liés…**, puis sélectionnez le test que vous souhaitez mettre en œuvre (**Wilcoxon, Signe, McNemar**), comme le montre la figure 2.16.

- **Test du signe**

  Le test du signe est relativement proche du test de Wicoxon, mais il est plus limité et par suite moins puissant. Il ne s'attache en effet qu'à une comparaison des signes des différences, sans procéder à un classement comme le fait le test de Wilcoxon.

Figure 2.16

Boîte de dialogue des tests de comparaison de deux échantillons appariés.



- **Test de McNemar**

  Le test de McNemar peut également être mis en œuvre dans le cas d'échantillons appariés, pour comparer les valeurs de deux variables dichotomiques (à deux dimensions).

Résumé

Première étape de l'analyse à proprement parler, la description des données permet de représenter les valeurs observées sur les différents individus de l'échantillon. L'**analyse univariée**, qui examine une seule variable à la fois, repose sur la **description** (fréquences, tendance centrale, dispersion, distribution) et la visualisation graphique des variables, ainsi que sur l'**inférence**, c'est-à-dire la comparaison à des valeurs déterminées. L'**analyse bivariée** permet d'aller plus loin par l'étude des relations entre deux variables, grâce aux **tris croisés** et aux principaux tests d'analyse bivariée : **tests d'association** (khi-deux) et **tests de comparaison** (test *t*, test K-S, test *U* de Mann-Whitney, etc.). Pour aller encore plus loin dans l'analyse, le chargé d'étude devra mettre en place des analyses multivariées, lesquelles seront abordées dans les chapitres suivants.

Pour aller plus loin

Evrard Y., Pras B., Roux E., *Market. Études et recherche en marketing,* Nathan, Paris, 2003.

Hair J. F., Anderson R. E., Tatham R. L., Black W. C., *Multivariate Data Analysis*, Prentice Hall International, New Jersey, 2007.

Malhotra N., Decaudin J. M., Bouguerra A., *Études marketing avec SPSS*, 5e éd., Pearson Education, Paris, 2007.

Tenenhaus M., *Méthodes statistiques en gestion*, Dunod, Paris, 2006.

Exercices

Exercice 1 Les tests

Énoncé

Répondez aux questions suivantes.

1. Quel(s) test(s) recommanderiez-vous à un chargé d'étude souhaitant comparer l'intention d'achat d'un produit avant et après son exposition dans un film publicitaire ?
2. Une compagnie de téléphonie mobile cherche à déterminer les principaux facteurs explicatifs de l'attrition, c'est-à-dire de la résiliation de l'abonnement en faveur d'un concurrent. En complément des données dont l'entreprise disposait dans sa base de données, une étude par téléphone a été commandée auprès d'un institut pour interroger les clients et les anciens clients. Interprétez les résultats mentionnés dans le tableau suivant.

……Item	Clients	Anciens clients	Signification
Âge moyen	47,6 ans	22,1 ans	,000
Durée de l'abonnement	7,1 ans	1,3 ans	,000
Possession d'un abonnement fixe	87 %	85 %	,372
Possession d'un abonnement Internet	72 %	79 %	,540
Possession d'un deuxième téléphone portable	13 %	23 %	,025
Degré de satisfaction* exprimé : – qualité globale du service	5,5	4,9	,459
– couverture du réseau	6,1	5,8	,248
– qualité des communications	5,5	2,3	,031
– qualité du centre d'appel	6,3	5,9	,462
– options de l'abonnement	5,7	3,2	,001
– nombre de SMS dans l'abonnement	5,8	5,2	,659
– prix de l'abonnement	6,1	4,0	,001
– coût total mensuel de l'abonnement	5,2	4,8	,001

* Mesuré sur une échelle de Likert en 7 points.

3. Quel test pourriez-vous mettre en place pour en apprendre davantage sur les deux derniers items du tableau ci-dessus : « prix de l'abonnement » et « coût total mensuel de l'abonnement » ?

Solution

1. Dans ce cas de figure, le chargé d'étude doit comparer la moyenne des réponses de deux échantillons à deux périodes distinctes, en d'autres termes avant et après l'exposition du produit dans un message publicitaire. Il s'agit donc d'une mesure sur échantillons appariés. Il pourra, par exemple, mettre en œuvre une extension du test *t*.

2. Les résultats de l'enquête comportent 7 résultats significatifs à un degré de confiance de 95 %. On peut donc conclure que les abonnés ayant préféré la concurrence sont en général plus jeunes et que leur abonnement était plus récent que celui des abonnés interrogés. En outre, ils sont plus nombreux à posséder un second téléphone portable et sont relativement moins satisfaits de la qualité des communications, des options de leur abonnement ainsi que du prix et du coût global mensuel de l'abonnement que les abonnés interrogés.

3. On peut réaliser un tri croisé des deux variables qualitatives et procéder à un test de khi-deux pour savoir si les deux variables sont liées. Comme nous n'avons aucune information sur le nombre d'observations, nous ne sommes pas certains de pouvoir respecter la condition de 5 observations par case du tableau. En outre, il est possible de procéder à un test sur échantillons appariés pour comparer les réponses aux deux questions : un test de Wicoxon serait alors approprié.

Exercice 2 Applications SPSS : l'enquête « point de vente » 2

Énoncé

Reprenons l'enquête sur le point de vente abordée dans la partie théorique de ce chapitre. Notre chargé d'étude cherche à en savoir davantage sur les données dont il dispose. Afin de progresser dans la maîtrise de l'outil SPSS, ouvrez le fichier « pointdevente.sav » disponible sur le site de l'ouvrage, et accompagnez le chargé d'étude dans sa réflexion en répondant aux questions suivantes.

1. Nous souhaitons en savoir un peu plus sur les répondants à l'enquête. Vous devez par conséquent poursuivre la description des variables de l'enquête que nous avons amorcée. Que pouvez-vous dire à propos des variables suivantes :
   a. progradio ?
   b. édition TV ?
   c. rubrikpress ?
2. Quel est le profil type du client de ce point de vente ? Que pouvez-vous en conclure sur le type de magasin dont il s'agit ?
3. L'enseigne mise sur ses clients les plus fidèles. L'équipe du magasin considère en effet que les clients ayant la plus forte intention d'effectuer leurs achats dans le magasin sont également ceux qui sont susceptibles de dépenser le plus. L'équipe a-t-elle raison de penser de la sorte ? Combien ces clients sont-ils susceptibles de dépenser pour un tee-shirt ? Les prix moyens affichés dans le magasin étant de 9 €, qu'en concluez-vous ?
4. L'enquête s'intéresse également aux goûts des clients potentiels. Nous avons lancé des pistes en ce qui concerne les prix, mais pouvez-vous aider l'équipe marketing du magasin à choisir la bonne musique d'ambiance : plutôt rock ou plutôt classique ?

Solution

1. Ces trois variables sont des variables qualitatives (nominales) et nous souhaitons les décrire. Il faut donc appeler la boîte de dialogue **Effectifs** dans le menu **Analyse**, puis le sous-menu **Statistiques descriptives…** Nous ne représenterons ici que la variable *progradio*, qui correspond à la question : « Quel type de programme radio écoutez-vous le plus souvent ? » et qui peut être décrite de la manière suivante (voir figure 2.17).

Figure 2.17

Effectifs de la variable *progradio*.

Statistiques

Quel type de programme de radio écoutez-vous le plus souvent?

N	Valide	385
	Manquante	15
Minimum		1

Quel type de programme de radio écoutez-vous le plus souvent?

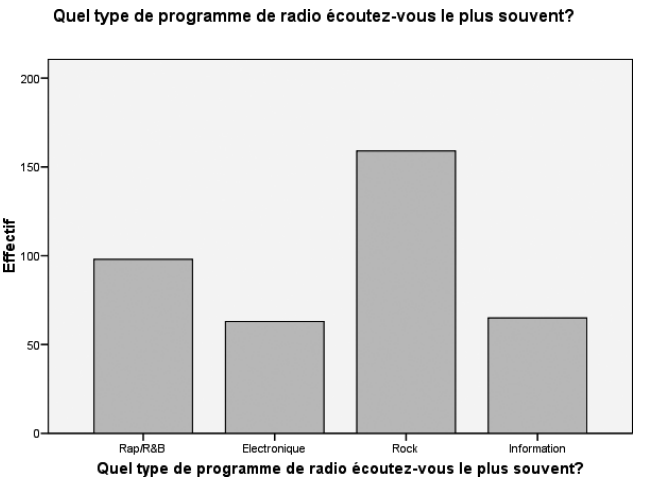
		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Rap/R&B	98	24,5	25,5	25,5
	Electronique	63	15,8	16,4	41,8
	Rock	159	39,8	41,3	83,1
	Information	65	16,3	16,9	100,0
	Total	385	96,3	100,0	
Manquante	Système manquant	15	3,8		
Total		400	100,0		

Nous avons choisi de représenter la variable *progradio* d'une façon relativement simple, en ne demandant que les effectifs, les pourcentages ainsi que le mode. Le mode représentant la valeur la plus fréquemment obtenue pour chaque modalité, le résultat est confirmé dans le tableau ci-dessus où la radio rock est la plus fréquemment écoutée (39,8 % des réponses).

On peut également représenter la variable par un diagramme bâtons (voir figure 2.18).

Figure 2.18

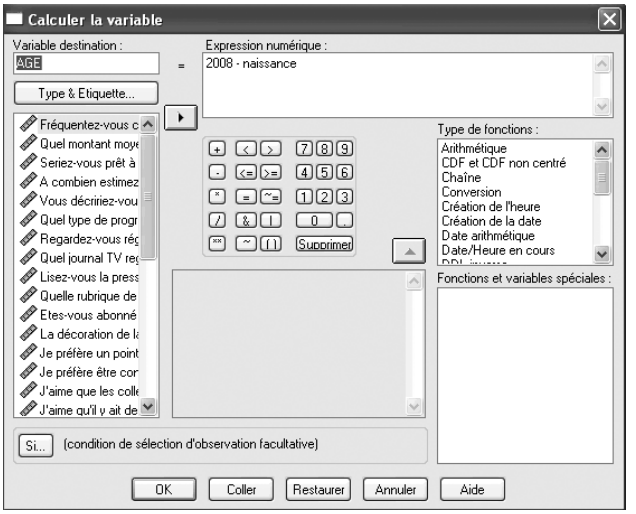
Diagramme bâtons de la variable *progradio*.



2. Pour établir le profil type du client de ce point de vente, il est nécessaire de décrire un certain nombre de variables de catégorisation, tels l'âge, le sexe (nous avons déjà décrit cette variable dans la partie théorique du chapitre), les revenus, le niveau d'études, etc. Dans le jeu de données, seule l'année de naissance est disponible. Il faut donc transformer cette variable afin de définir l'âge des répondants. Dans le menu **Transformer**, ouvrez la boîte de dialogue **Calculer la variable**. Pour calculer l'âge des répondants, il suffit de retirer l'âge de chaque répondant à l'année actuelle (2008) comme le montre la figure 2.19.

Figure 2.19

Boîte de dialogue Calculer une variable.



Nous appelons AGE la nouvelle variable créée. On peut maintenant calculer l'âge moyen des répondants (voir figure 2.20).

Figure 2.20

Âge des répondants.

Statistiques descriptives

	N	Minimum	Maximum	Moyenne	Ecart type
AGE	400	33,00	80,00	49,3375	9,96830
N valide (listwise)	400				

Décrivons maintenant les revenus ainsi que le niveau d'études des répondants (voir figures 2.21 et 2.22).

Figure 2.21

Revenus des répondants.

Quels sont approximativement les revenus de votre foyer?

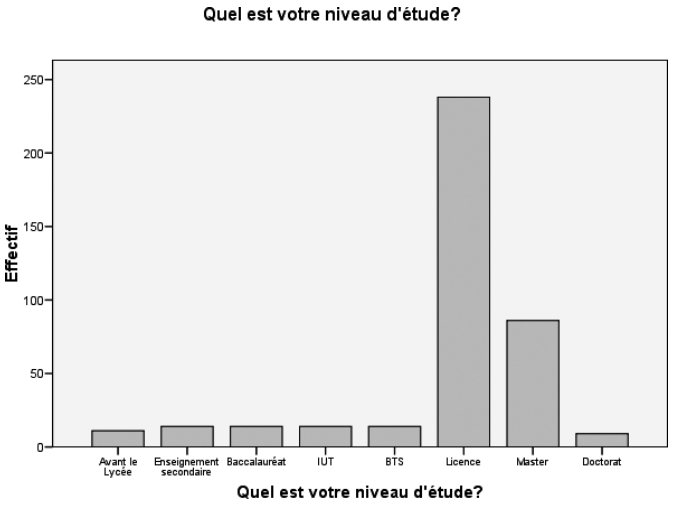
		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	<15 000€	26	6,5	6,5	6,5
	15 000€ - 24 999€	34	8,5	8,5	15,0
	25 000€ - 49 999€	56	14,0	14,0	29,0
	50 000€ - 74 999€	99	24,8	24,8	53,8
	75 000€ - 99 999€	65	16,3	16,3	70,0
	100 000 - 149 999€	54	13,5	13,5	83,5
	+ de 150 000€	66	16,5	16,5	100,0
	Total	400	100,0	100,0	

Pour conclure rapidement, on peut dire que l'âge moyen du répondant est de 39,33 ans, qu'il s'agit de ménages aisés (seuls 29 % des foyers gagnent moins de 50 000 € annuels), ayant fait des études supérieures (plus de 80 % ont au moins une licence). Si l'on complète en incorporant les éléments vus dans la partie cours, on peut également dire qu'il s'agit aussi bien d'hommes que de femmes, et que le montant moyen mensuel dépensé dans le magasin est relativement élevé (pour en savoir plus, il faudrait mettre en place une analyse typologique). Il pourrait s'agir d'une enseigne de prêt-à-porter moyen de gamme, même si les données dont nous disposons sont relativement limitées pour ce genre de conclusion.

3. Pour apporter une réponse à l'équipe marketing du magasin, il faut d'abord sélectionner les répondants qui nous intéressent. Allez dans le menu **Données > Sélectionner des observations…** Sélectionnez les observations de la variable *intention* (« Seriez-vous prêt à faire vos achats dans ce point de vente? ») **selon la condition logique : intention = 5**

Figure 2.22

Desription du niveau d'études des répondants.



(5 étant le score de la plus haute intention d'achat). Une fois que vous avez cliqué sur **OK**, les autres observations sont barrées dans l'éditeur de données. Nous cherchons donc à calculer la moyenne des dépenses du groupe des répondants ayant la plus forte intention d'achat, en essayant d'établir s'ils sont prêts à payer plus que la moyenne des clients du magasin (on suppose ici que le prix moyen est le prix affiché). Il s'agit d'un test *t* sur échantillon unique où la valeur comparée sera le prix affiché (p = 9). Les résultats apparaissent dans les tableaux de la figure 2.23.

Figure 2.23

Comparaison de moyenne de la variable prix.

Statistiques sur échantillon unique

	N	Moyenne	Ecart-type	Erreur standard moyenne
A combien estimez-vous le prix moyen d'un t-shirt dans ce point de vente?	79	18.1266	6.90664	.77706

Test sur échantillon unique

	Valeur du test = 9					
	t	ddl	Sig. (bilatérale)	Différence moyenne	Intervalle de confiance 95% de la différence	
					Inférieure	Supérieure
A combien estimez-vous le prix moyen d'un t-shirt dans ce point de vente?	11,745	78	,000	9.12658	7.5796	10.6736

Les clients ayant la plus forte intention d'achat dépensent en moyenne plus de 18 € pour un tee-shirt, soit plus du double du prix affiché. Ces résultats sont significatifs ($p < 0.05$), ce qui signifie que la différence moyenne constatée (+9,13 €) est statistiquement différente du prix moyen affiché. Une piste pour élargir la fourchette des prix pratiqués ?

4. On peut comparer les réponses à deux questions mesurées de la même manière par le biais d'un test *t* pour échantillons appariés, ou bien en mettant en place un test de Wilcoxon si l'on préfère un test non paramétrique. Attention ! Vous devez sélectionner de nouveau l'ensemble des répondants. Les tableaux de la figure 2.24 présentent les statistiques et les résultats du test *t*.

Figure 2.24

Statistiques et test *t* sur échantillons appariés.

Statistiques pour échantillons appariés

		Moyenne	N	Ecart-type	Erreur standard moyenne
Paire 1	J'aime une musique d'ambiance classique	2,36	400	1,451	,073
	J'aime une musique d'ambiance rock	3,35	400	1,239	,062

Corrélations pour échantillons appariés

		N	Corrélation	Sig.
Paire 1	J'aime une musique d'ambiance classique & J'aime une musique d'ambiance rock	400	-,403	,000

Test échantillons appariés

		Différences appariées					t	ddl	Sig. (bilatérale)
		Moyenne	Ecart-type	Erreur standard moyenne	Intervalle de confiance 95% de la différence Inférieure	Supérieure			
Paire 1	J'aime une musique d'ambiance classique - J'aime une musique d'ambiance rock	-,990	2,257	,113	-1,212	-,768	-8,774	399	,000

Le premier tableau reprend les moyennes de réponses ainsi que les statistiques associées. On constate que la préférence va à la musique rock. Le second tableau permet de rejeter l'hypothèse nulle relative à l'égalité des deux mesures. Il existe donc une préférence significative pour une musique d'ambiance de type rock.

On peut également obtenir ces résultats en utilisant le test de Wilcoxon comme l'indiquent les résultats reportés sur la figure 2.25.

Figure 2.25

Rangs et test de Wilcoxon sur échantillons appariés.

Rangs

		N	Rang moyen	Somme des rangs
J'aime une musique d'ambiance rock - J'aime une musique d'ambiance classique	Rangs négatifs	118[a]	159,92	18871,00
	Rangs positifs	259[b]	202,25	52382,00
	Ex aequo	23[c]		
	Total	400		

a. J'aime une musique d'ambiance rock < J'aime une musique d'ambiance classique

b. J'aime une musique d'ambiance rock > J'aime une musique d'ambiance classique

c. J'aime une musique d'ambiance rock = J'aime une musique d'ambiance classique

Test[b]

	J'aime une musique d'ambiance rock - J'aime une musique d'ambiance classique
Z	-7,995[a]
Signification asymptotique (bilatérale)	,000

a. Basée sur les rangs négatifs.

b. Test de Wilcoxon

Le test de Wilcoxon confirme le résultat précédent. Le sens de la statistique confirme également le sens de la différence examinée, en faveur du second élément de la paire : la musique rock.

Chapitre 3

Simplifier les données

La simplification ou l'agrégation des données est fondamentale, elle sert à identifier les différentes dimensions d'un concept. L'analyse factorielle est utilisée pour décrire les données en un nombre agrégé de facteurs. Elle traduit une matrice de nombres difficile à lire par une série de tableaux plus simples, représentés sous forme de graphiques.

Les principes de validation d'une échelle de mesure : les notions de validité et de fiabilité sont tout d'abord exposés. Ensuite, nous présentons l'analyse factorielle et ses applications.

1 Principes de validation d'une échelle de mesure

Nous avons vu dans le chapitre 1 que les concepts étaient mesurés avec plusieurs questions ou items. Par exemple, pour estimer l'attitude du client à l'égard d'un produit, le chargé d'étude pose des questions qui permettent de bien saisir les différentes facettes de ce concept (part affective, cognitive, etc..). Ensuite, il faut vérifier que ces différentes questions ou items mesurent bien ce que l'on cherche à mesurer, afin d'obtenir au final des résultats plus proches de la réalité.

Dans l'article intitulé « Un paradigme pour développer de meilleures mesures des construits marketing », Churchill (1979) propose une procédure pour renforcer la validité et la fiabilité des mesures. Après avoir sélectionné des échelles (jeu d'items pour mesurer un concept), il s'agit, dans un premier temps, de les soumettre à l'analyse factorielle exploratoire puis au test de la fiabilité [1].

1.1 La validité d'une échelle de mesure

La validité d'une échelle de mesure désigne sa capacité à appréhender un phénomène (Hair *et al.*, 2006). Les tests de validité ont pour objectif de vérifier si les différents items d'un instrument sont une bonne représentation du phénomène étudié : mesure-t-on ce que l'on cherche à mesurer ? (Evrard *et al.*, 2003).

La validité prend plusieurs formes ; il existe donc plusieurs techniques pour la vérifier :

- **la validité faciale ou de contenu** : il s'agit de savoir si la mesure capture les différents aspects du phénomène étudié. Elle est fondée sur le jugement du chercheur et de ses pairs. Par exemple, lors du test du questionnaire, des experts du domaine peuvent émettre un avis sur la capacité des items à recouvrir tous les aspects d'un concept ;
- **la validité de trait ou de construit** : est-ce que les différents indicateurs offrent une bonne représentation du phénomène étudié ? Il faut vérifier si les indicateurs censés mesurer le même phénomène sont corrélés (validité convergente) et s'ils se distinguent des indicateurs censés mesurer des phénomènes différents (validité discriminante) (Evrard *et al.*, 2003) :
  - **la validité convergente** est établie lorsque les mesures d'un même construit sont corrélées ;
  - **la validité discriminante** est destinée à s'assurer que les indicateurs de mesure d'un construit sont faiblement corrélés aux indicateurs de mesure d'autres construits, conceptuellement distincts du premier. L'analyse factorielle exploratoire (AFE) permet de tester ces deux validités ;
- **la validité nomologique ou prédictive** résulte de la conformité des relations entre les mesures d'un concept et celles d'autres concepts avec les prédictions de la théorie (Evrard *et al.*, 2003). Cette étape de validation intervient au cours de la phase confirmatoire.

1. Puis, dans une phase de validation, les échelles modifiées après suppressions d'énoncés subissent une deuxième fois ces procédures, on parle d'analyse confirmatoire. Cette seconde étape vise à connaître les qualités psychométriques des instruments de mesure.

1.2 La fiabilité d'une échelle de mesure

La fiabilité correspond au degré avec lequel les instruments utilisés mesurent de façon constante le construit étudié (Evrard *et al.*, 2003). Par conséquent, une échelle est fidèle si l'on retrouve plusieurs fois les mêmes résultats sur les mêmes sujets. Trois méthodes permettent de tester la fiabilité d'une mesure :

- **la méthode du « *test/retest* »** : le questionnaire est administré deux fois à la même population et les résultats obtenus sont comparés. Cette technique est particulièrement appropriée pour la mise au point d'instrument de mesure ;
- **la méthode du « *Split half* », ou des deux moitiés** : le questionnaire est administré au même moment à des échantillons différents (l'échantillon est scindé en deux) et les résultats sont comparés. Il existe cependant un risque de sélection ; les deux échantillons sont-ils appariés ? se ressemblent-ils ?
- **la technique des formes alternatives** : il s'agit d'introduire dans le questionnaire plusieurs questions sur le même phénomène mais formulées différemment. Le questionnaire est administré aux mêmes individus. Le coefficient alpha de Cronbach est calculé pour vérifier si les énoncés partagent des notions communes, et s'ils sont en cohérence entre eux.

Focus 3.1 Estimer la fiabilité avec le coefficient alpha de Cronbach

L'alpha de Cronbach est un coefficient de fiabilité qui mesure la cohérence interne d'une échelle construite à partir d'un ensemble d'items. La pratique consiste à réduire un grand nombre d'items initiaux dans un processus itératif de conservation/élimination des items en fonction de la valeur du coefficient alpha, qui varie entre 0 et 1. Plus la valeur de l'alpha est proche de 1, plus la cohérence interne de l'échelle (sa fiabilité) est forte. On élimine donc les items qui diminuent le score, et on conserve ceux qui contribuent à augmenter l'alpha. L'examen de l'alpha de Cronbach évite au chargé d'étude de tomber dans un travers fréquent qui consiste à reprendre un questionnaire existant sans se préoccuper de ses problèmes de mesure.

Le seuil d'acceptabilité de l'alpha varie selon l'objectif de la recherche. Pour une étude exploratoire, un coefficient plus faible est acceptable (0,7) alors que dans le cadre d'une recherche fondamentale, il doit être plus élevé (> 0,8) (Nunnally et Bernstein, 1994). Lorsqu'une échelle est utilisée pour comparer des groupes, un alpha de 0,8 est satisfaisant, et il est inutile d'essayer d'obtenir un niveau supérieur. De Vellis (2003) propose une typologie (voir tableau 3.1).

Tableau 3.1 : Les valeurs de l'alpha de Cronbach

< 0,6	Insuffisant
entre 0,6 et 0,65	Faible
entre 0,65 et 0,7	Minimum acceptable
entre 0,7 et 0,8	Bon
entre 0,8 et 0,9	Très bon
> 0,9	Considérer la réduction du nombre d'items

Il existe une relation entre le nombre d'items et la valeur de l'alpha : un nombre restreint d'items (de deux à trois) donne un alpha généralement plus faible (0,6) qu'une mesure de quatre énoncés (0,7). Au-delà de 0,9, l'alpha risque, en revanche, de traduire davantage une redondance inter-items, appauvrissant ainsi le domaine conceptuel étudié (Peterson, 1995). Il est, par conséquent, recommandé de ne pas dépasser le seuil de 0,9.

Le logiciel SPSS fournit les niveaux du coefficient d'alpha de l'échelle lorsque chaque item est supprimé. Les items dont la suppression améliore sensiblement le coefficient ne sont généralement pas retenus si la validité de contenu ne s'en trouve pas amoindrie.

Focus 3.2 Le traitement des items inversés

La conception d'un questionnaire demande des précautions (plusieurs items, non-réponse possible, clarté de la question, ordre des questions, etc.) car la formulation des questions peut influencer la mesure d'un concept. L'inversion d'item est souvent employée pour s'assurer de la validité et de la fiabilité de la mesure (par exemple, la satisfaction à l'égard d'un service est mesurée par un item : « je suis pleinement satisfait par ce service » et un autre, inversé, « ce service ne me satisfait pas pleinement »).

Nous cherchons à évaluer dans quelle mesure les items utilisés sont de bons indicateurs des concepts qu'ils sont censés mesurer. Pour cela, il est généralement conseillé de réaliser une analyse factorielle exploratoire pour vérifier que les items se *« regroupent »* bien de la manière prévue, et de calculer ensuite le coefficient alpha de Cronbach qui évalue la fiabilité de chaque échelle (Churchill, 1979).

2 L'analyse factorielle

L'analyse factorielle est une méthode exploratoire d'analyse des tableaux de contingence développée essentiellement par J.-P. Benzecri durant la période 1970-1990. Elle désigne un ensemble de méthodes statistiques multivariées dont le principal objectif est de définir la structure des corrélations entre un grand nombre de variables (par exemple, les réponses à un questionnaire) en déterminant un ensemble de dimensions communes appelés facteurs.

2.1 Les utilisations de l'analyse factorielle

L'analyse factorielle sert à identifier les **dimensions** de la structure et à déterminer dans quelle mesure chaque variable peut expliquer chaque dimension.

Les deux objectifs de l'analyse factorielle sont :

- **Résumer les données.** L'analyse factorielle fait ressortir les dimensions sous-jacentes qui, une fois interprétées, décrivent les données de manière synthétique.
- **Réduire les données.** Elle calcule des scores pour chaque dimension et les substitue aux variables originelles.

Alors que dans les autres méthodes (régressions, analyse de variance, etc.) les variables sont considérées comme des variables soit dépendantes, soit indépendantes, dans l'analyse factorielle, toutes les variables sont considérées chacune par rapport aux autres. Les **facteurs** sont formés pour maximiser l'explication de l'ensemble des variables et non pour prédire des variables dépendantes. Dès lors, l'analyse factorielle est appropriée dans une optique exploratoire (analyse factorielle exploratoire ou AFE).

EXEMPLE

Les critères importants dans l'évaluation d'un club de sport

Dans une enquête sur les attentes des clients vis-à-vis de leur salle de sport, on interroge les individus sur une vingtaine de critères. L'analyse factorielle sert à regrouper les attentes en trois ou quatre points plus simples. Elle agrège les variables en facteurs ou combinaisons de variables. L'objectif est de rendre l'information plus synthétique et facile à lire sur une carte factorielle (voir tableaux 3.2 et 3.3).

Tableau 3.2 : Exemple d'application de l'analyse factorielle

	Rencontre	Muscles	Esthétisme	Défoulement	Santé	Dynamisme	Prise en charge	Confort	Économie	Lieu agréable
1	4	1	4	2	4	1	1	2	1	2
2	1	2	4	5	4	1	1	1	1	1
3	2	4	2	4	3	1	1	2	4	2
4	3	4	2	4	3	3	3	2	1	2
5	1	4	3	4	4	4	4	3	2	3
6										

À titre d'exemple, le confort, les aspects défoulement, dynamisme et santé représentent peut-être en fait la même chose : être en forme.

Tableau 3.3 : Exemple d'application de l'analyse factorielle (*suite*)

	Facteur 1 Forme	Facteur 2 Contact	Facteur 3
1			
2			
3			
4			
5			
6			

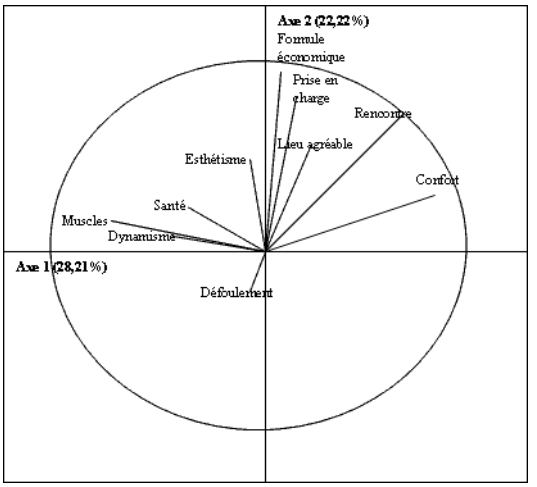
La solution de l'analyse factorielle est trouvée par essai/erreur et le jugement s'établit en fonction des concepts (voir figure 3.1). Sur l'axe horizontal de la figure, à gauche les atten-

tes des clients portent sur la forme physique ; à droite, sur le confort de la salle. Sur l'axe vertical s'opposent le côté sociable du club de sport et le besoin de s'y défouler.

Au total, la variance restituée par ces deux axes (les deux premiers facteurs) est de 50,43.

Figure 3.1

Représentation graphique de l'analyse factorielle.



L'analyse factorielle exploratoire permet d'identifier des groupes d'items qui covarient les uns avec les autres et semblent représenter des variables latentes pertinentes. Autrement dit, l'AFE consiste à explorer la relation entre des variables mesurées, afin de déterminer si ces relations peuvent être résumées par un nombre moins important de construits latents.

L'AFE permet de vérifier le nombre de dimensions ou, plus souvent, l'unidimensionalité d'un concept. En effet, un concept peut comporter une ou plusieurs facettes. Par exemple, l'implication comporte une composante affective, une composante calculée et une composante normative. Lorsque l'on fait appel à des échelles de mesure déjà utilisées, l'AFE permet de vérifier si l'on retrouve, pour l'échantillon étudié, la même structure factorielle. Elle fera alors ressortir autant de facteurs que le construit a de dimensions (un seul facteur si le construit est unidimensionnel). Dans le cadre du développement de nouveaux instruments, l'AFE permet de constater si les items correspondent effectivement aux concepts présentés aux répondants.

2.2 Les conditions et options de l'analyse factorielle

En fonction des caractéristiques de l'échantillon et des données collectées, plusieurs options sont possibles pour la réalisation d'une AFE (analyse factorielle exploratoire). Nous verrons, dans un premier temps, la taille de l'échantillon requise avant de présenter les différentes options et tests permettant de valider les résultats d'une AFE.

La taille de l'échantillon nécessaire

La taille de l'échantillon dépend du nombre d'items soumis à l'AFE. Il faut un minimum de cinq observations par item (un ratio de 10 pour 1 est préférable). Le nombre total d'observations doit être d'au moins 50 et il est souhaitable d'interroger au moins 100 individus.

La méthode d'extraction utilisée

La méthode d'extraction la plus employée est l'analyse en composantes principales (ACP). L'ACP a pour objet de résumer l'ensemble des données quantitatives d'un tableau individus/variables. En effet, l'ACP synthétise les données en construisant un petit nombre de variables nouvelles, les composantes principales. Les éléments critiques de la grille peuvent alors être captés rapidement, à l'aide de représentations graphiques établies à partir des ACP.

Le choix de la matrice des données

Il est possible de travailler sur la matrice de corrélation ou sur la matrice de covariance. Pour simplifier, ce choix s'effectue ainsi :

- **matrice de corrélation :** lorsque les variables sont mesurées avec des échelles différentes ;
- **matrice de covariance :** lorsque l'on applique l'analyse factorielle à plusieurs groupes avec des variances différentes pour chaque variable.

L'adéquation des données

Avant de réaliser l'analyse, il est important de s'assurer que les données sont factorisables. Elles doivent former un ensemble cohérent pour pouvoir y chercher des dimensions communes qui aient un sens et qui ne soient pas des artefacts statistiques (Evrard *et al.*, 2003). La matrice des données doit comporter suffisamment de corrélations pour justifier la réalisation d'une AFE. Plusieurs indicateurs peuvent être utilisés :

- **La matrice des corrélations anti-image** représente la valeur négative des corrélations partielles. Des corrélations anti-image importantes indiquent que la matrice des données n'est peut-être pas adaptée à l'AFE.
- **Le test de Sphéricité de Bartlett** examine la matrice des corrélations dans son intégralité et fournit la probabilité de l'hypothèse nulle selon laquelle toutes les corrélations sont de zéro.
- **La *« Measure of Sampling Adequacy »* (MSA) ou *Kaiser-Meyer-Olkin* (KMO)** indique dans quelle proportion les variables retenues forment un ensemble cohérent et mesurent de manière adéquate un concept. Elle teste si les corrélations partielles entre les variables sont faibles.

Des valeurs de KMO comprises entre 0,3 et 0,7 représentent des solutions factorielles acceptables. Ce test, d'abord réalisé pour chaque variable, doit ensuite être repris avec l'ensemble des variables (Hair *et al.*, 2006).

L'extraction des facteurs

Il n'existe pas de base quantitative exacte pour déterminer le nombre de facteurs à extraire. Les critères sont souvent choisis sur la part de variance de chaque item qu'un facteur permet d'expliquer :

- **l'*« eigenvalue »*, ou règle des valeurs propres > 1 ou règle de Kaiser-Guttman** : une valeur propre représente la quantité d'informations capturée par un facteur. Un facteur qui aurait une valeur propre inférieure à 1 représenterait moins d'informations qu'un simple item.
- **le *« Scree Test »*, ou test du coude ou de l'éboulis** : ce test se fonde également sur les valeurs propres des facteurs mais dans une perspective relative et non absolue. Étant

donné que chaque facteur est extrait d'une matrice qui est le résidu de l'extraction précédente, la quantité d'informations contenue dans les facteurs successifs décroît. Lorsque, entre deux facteurs, la décroissance en termes d'informations devient faible ou nulle, on peut estimer que le dernier facteur ne contient pas suffisamment d'informations pour être retenu.

- **le critère du pourcentage de variance** : il s'agit d'une approche par laquelle on observe les pourcentages cumulés de la variance extraite par les facteurs successifs. L'objectif est de s'assurer qu'un facteur explique une quantité significative de variance. Il est souvent conseillé d'arrêter l'extraction lorsque 60 % de la variance expliquée est extraite (Hair *et al.*, 2006).

La rotation des facteurs

Afin de pouvoir interpréter les facteurs, il est généralement nécessaire de réaliser une rotation. Celle-ci permet d'identifier des groupes de variables fortement liés les uns aux autres. La rotation fait en sorte que chaque item ne soit fortement lié qu'à un seul facteur. Cette opération est réalisée par une redistribution de la variance des premiers facteurs extraits aux facteurs successifs, afin d'aboutir à une structure factorielle plus simple (Hair *et al.*, 2006). Lorsque les axes sont maintenus à 90 degrés, on parle de rotation orthogonale ; lorsque les axes ne sont pas contraints à être indépendants, on parle de rotation oblique.

Il existe plusieurs méthodes de rotation :

- **Varimax** : rotation orthogonale qui minimise le nombre de variables ayant de fortes corrélations sur chaque facteur. Simplifie l'interprétation des facteurs.
- **Oblimin direct** : rotation oblique, c'est-à-dire dans laquelle les axes se positionnent en fonction des items et ne sont donc pas orthogonaux.
- **Quartimax** : méthode qui minimise le nombre de facteurs requis pour expliquer chaque variable. Simplifie l'interprétation des variables observées.
- **Equamax** : méthode de rotation qui minimise à la fois le nombre de variables qui pèsent fortement sur un facteur et le nombre de facteurs requis pour expliquer une variable (combinaison des méthodes Varimax et Quartimax).

Focus 3.3 L'analyse factorielle exploratoire : rotation orthogonale ou oblique ?

Les critères de choix entre la **rotation orthogonale** (Varimax) et la **rotation oblique** sont les suivants :

La rotation orthogonale maintient les axes de l'espace factoriel en angle droit. Ce type de rotation permet de minimiser le nombre d'items ayant des contributions élevées sur un axe et donc de simplifier les facteurs. Elle permet d'obtenir une structure factorielle plus claire.

Si la corrélation entre facteurs est faible, inférieure à 0,15 (De Vellis, 2003) ou à 0,3 (Nunnally et Bernstein, 1994), la rotation orthogonale sera préférée pour sa simplicité. Toutefois, si l'on a des raisons de penser que des items ou facteurs sont corrélés, il est logique de réaliser une rotation oblique. On peut également comparer la solution avec rotation oblique et rotation orthogonale. S'il est possible d'assigner un item au même facteur dans les deux cas, alors la rotation orthogonale sera choisie pour sa simplicité.

Dans la grande majorité des cas, une rotation orthogonale est suffisante pour aboutir à une structure simple. Hair *et al.* (2006) estiment cependant que la rotation oblique est conseillée si l'on souhaite déterminer des facteurs représentant des concepts qui seront analysés postérieurement car la structure factorielle obtenue possède une plus grande stabilité.

L'AFE pour vérifier le nombre de dimensions d'un concept

L'analyse factorielle est utilisée pour vérifier **la validité de trait ou de construit.** Il s'agit de tester et de purifier les échelles d'un questionnaire. L'AFE permet de s'assurer que l'échelle évalue précisément et exclusivement le construit qu'elle est censée mesurer. Lorsque le construit est unidimensionnel, l'AFE fera ressortir un seul facteur, plusieurs pour **les construits multidimensionnels**. Il est aussi possible de fixer *a priori* le nombre de facteurs sous SPSS.

Nous traitons ici de la question des items et facteurs à retenir ou, au contraire, à supprimer, suite à une AFE. En effet, lorsque les facteurs sont extraits, il est nécessaire d'évaluer la validité convergente et discriminante au niveau de l'item ainsi que la fiabilité des échelles. La validité convergente concerne le fait que les réponses obtenues par différents indicateurs du même construit soient fortement corrélées; la validité discriminante est démontrée lorsque la mesure d'un construit déterminé est faiblement corrélée à une mesure d'un autre construit.

Ces analyses sont réalisées pour chaque échelle. Les items présumés mesurer un même construit doivent donc être fortement corrélés les uns aux autres (validité convergente) et faiblement corrélés aux items censés mesurer d'autres construits (validité discriminante). Le niveau du coefficient structurel de l'AFE (composante) sert à déterminer si l'item satisfait au critère de validité convergente. Le tableau 3.4 présente les niveaux de significativité des contributions factorielles des items selon la taille de l'échantillon étudié.

Tableau 3.4 : Niveau de significativité des coefficients structurels selon la taille de l'échantillon

Niveau des coefficients structurels	Taille de l'échantillon nécessaire
0,30	350
0,35	250
0,40	200
0,45	150
0,50	120
0,55	100
0,60	85
0,65	70
0,70	60
0,75	50

Source : adapté de Hair *et al.*, 2006.

L'épuration d'une échelle de mesure se fait en deux temps :

D'une part, pour **les coefficients structurels ou composantes,** un seuil est déterminé en fonction de la taille de l'échantillon. Par exemple, pour un test d'échelle sur un échantillon de 200 individus, un seuil de 0,40 sera retenu. Pour les échelles multidimensionnelles, sont éliminés les items dont les poids factoriels sont supérieurs à 0,30 sur plusieurs facteurs et ceux n'ayant aucune contribution supérieure ou égale à 0,50 sur l'une des composantes principales identifiées. Ces seuils peuvent aussi varier en fonction de la taille de l'échantillon (Hair *et al.*, 2006).

D'autre part, la formation des facteurs repose sur l'importance des variables initiales sur ces facteurs. **Les « communalités »** (part de variance expliquée par l'item) doivent dépasser 0,5 et si possible 0,7. Le niveau de représentation est considéré comme moyen pour un seuil de 0,40, bon pour un seuil de 0,65 et excellent lorsque la communalité dépasse 0,80 (Evrard *et al.*, 2003).

SPSS

Dans cet exemple, nous testons l'échelle destinée à mesurer l'ambition professionnelle. Cette échelle unidimensionnelle de 10 items est issue de la littérature. Les réponses aux questions sont collectées grâce à une échelle de Likert à cinq échelons allant de « Pas du tout d'accord » à « Tout à fait d'accord » (voir tableau 3.5).

Tableau 3.5 : Exemple de l'échelle destinée à mesurer l'ambition

Item 1 - J'aimerais avoir un poste plus important et que les autres m'envient.
Item 2 - J'aime bien discuter avec des gens importants.
Item 3 - Je veux être une personne importante dans la communauté.
Item 4 - J'admire beaucoup les gens qui ont gravi les échelons et sont au sommet.
Item 5r[1] - Si j'avais suffisamment d'argent, je ne travaillerais plus*.
Item 6 - Même si je gagnais beaucoup d'argent au jeu, je continuerais à exercer mon métier.
Item 7r - Si je pouvais toucher le chômage, je préférerais ne pas travailler*.
Item 8 - J'aime être admiré(e) pour ma réussite.
Item 9r - Je n'aime pas être remarqué(e)*.
Item 10 - J'aime que des employés me demandent conseil.

1. Le r signifie que cet item est inversé.

Les 10 items sont, dans un premier temps, soumis à une analyse factorielle exploratoire (méthode de l'ACP), afin de vérifier la structure du construit mesuré. Cette échelle est testée avec un échantillon de 106 individus.

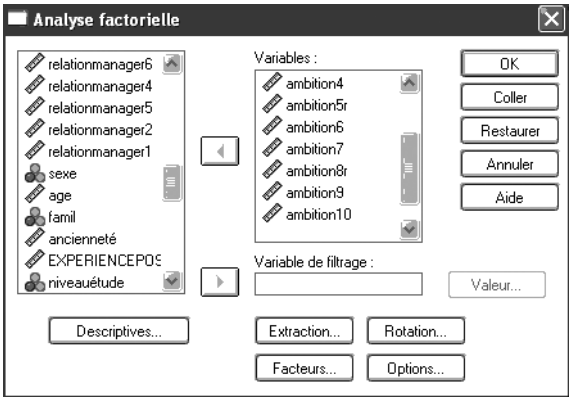
Ouvrez le fichier « challenge » [1]. Allez dans le menu **Analyse > Positionnement > Analyse factorielle.** Une boîte de dialogue apparaît (voir figure 3.2).

1. Vous trouverez ce fichier à l'adresse : http://www.pearsoneducation.fr.

Transférez les items destinés à mesurer l'ambition en les sélectionnant chacun à leur tour et en cliquant sur la flèche.

Figure 3.2

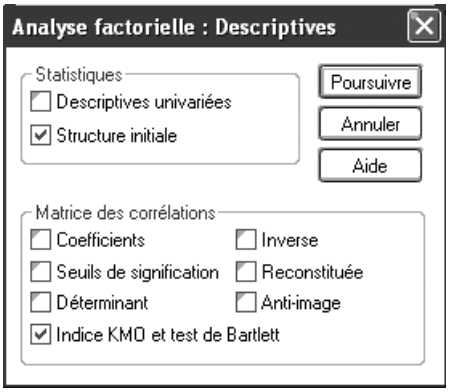
Demande d'analyse factorielle.



Avant de lancer l'AFE, plusieurs commandes sont à effectuer. Afin de vérifier l'adéquation des données, on peut demander l'indice KMO et le test de Bartlett par le bouton **Descriptives** dans la boîte de dialogue précédente. La structure initiale (précochée) donne les communalités, les valeurs propres et la part de variance expliquée initiale (voir figure 3.3).

Figure 3.3

Demande de l'indice KMO et du test de Bartlett.



Cliquez ensuite sur **Poursuivre** pour revenir à la boîte de dialogue initiale.

Pour sélectionner la méthode de l'analyse factorielle, cliquez sur **Extraction** et la boîte de dialogue de la figure 3.4 apparaît.

Nous pouvons choisir **la méthode de l'analyse factorielle** (composantes principales ; facteurs communs, etc.). Nous sélectionnons **Composantes principales**.

Pour obtenir les facteurs, le logiciel présélectionne les valeurs propres supérieures à 1. Mais il est aussi possible de déterminer le nombre de facteurs. Dans une optique exploratoire, nous laissons libre ce nombre de facteurs.

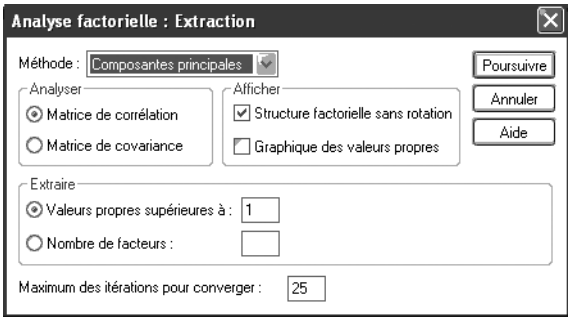
Le choix de la matrice de départ est aussi fixé dans cette boîte de dialogue : la matrice de corrélation est présélectionnée. Nous gardons cette matrice pour l'analyse.

On peut demander un graphique des valeurs propres qui sert à éliminer des facteurs avec le test du coude.

Cliquez ensuite sur **Poursuivre** pour revenir à la boîte de dialogue initiale.

Figure 3.4

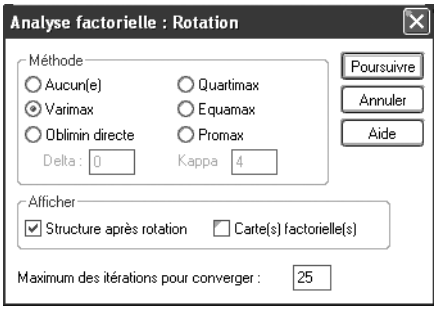
Choix de la méthode d'extraction, de la matrice de départ et demande de graphique.



Pour sélectionner la méthode de rotation, cliquez sur **Rotation** et la boîte de dialogue de la figure 3.5 apparaît. Cochez la méthode choisie, nous sélectionnons Varimax.

Figure 3.5

Choix de la méthode de rotation.



Cliquez ensuite sur **Poursuivre** pour revenir à la boîte de dialogue puis sur **OK** pour lancer l'AFE.

Les résultats de l'analyse apparaissent dans l'onglet résultats (voir figure 3.6).

Figure 3.6

Interprétation des résultats de l'AFE : KMO, test de Bartlett et communalités.

Indice KMO et test de Bartlett

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,816
Test de sphéricité de Bartlett	Khi-deux approximé	1041,747
	ddl	45
	Signification de Bartlett	,000

Qualité de représentation

	Initial	Extraction
ambition1	1,000	,474
ambition2	1,000	,453
ambition3r	1,000	,472
ambition4	1,000	,408
ambition5r	1,000	,745
ambition6	1,000	,727
ambition7	1,000	,537
ambition8r	1,000	,241
ambition9	1,000	,192
ambition10	1,000	,573

Méthode d'extraction : Analyse en composantes principales.

L'indice KMO (0,816) ainsi que le test de Bartlett permettent d'accepter les résultats de cette analyse factorielle.

Les résultats montrent que les deux facteurs qui n'expliquent pas plus de 50 % (48,21) de la variance (voir figure 3.7). Nous éliminons les items dont les communalités sont trop faibles (soit 8r et 9 qui ont des communalités respectives de 0,24 et 0,19).

Figure 3.7

Interprétation des résultats de l'AFE : pourcentage de variance expliquée, nombre de facteurs.

Variance totale expliquée

	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus			Somme des carrés des facteurs retenus pour la rotation		
Composante	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	3,413	34,127	34,127	3,413	34,127	34,127	3,251	32,511	32,511
2	1,408	14,084	48,210	1,408	14,084	48,210	1,570	15,699	48,210
3	,925	9,252	57,462						
4	,863	8,632	66,094						
5	,688	6,880	72,974						
6	,664	6,644	79,618						
7	,640	6,404	86,022						
8	,519	5,192	91,214						
9	,450	4,502	95,717						
10	,428	4,283	100,000						

Méthode d'extraction : Analyse en composantes principales.

En outre, la matrice des composantes atteste que deux items (5r et 6) ne se trouvent pas sur le même facteur (voir figure 3.8). Or, ce second facteur n'explique, à son tour, qu'une faible part de la variance. Nous éliminons donc ces deux items.

Figure 3.8

Interprétation des résultats de l'AFE : matrice des composantes (coefficients structurels).

Matrice des composantes après rotation[a]

	Composante	
	1	2
ambition1	,687	,040
ambition2	,665	,104
ambition3r	,656	,206
ambition4	,620	,152
ambition5r	,116	,855
ambition6	,054	,851
ambition7	,727	,092
ambition8r	,466	,152
ambition9	,436	-,048
ambition10	,755	-,056

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation de Kaiser.
a. La rotation a convergé en 3 itérations.

À ce stade, il faut refaire une AFE en rappelant la boîte de dialogue ou en allant dans le menu **Analyse > Factorisation > Analyse factorielle**. La boîte de dialogue de la figure 3.9 apparaît.

Il faut alors faire passer les items éliminés (5r, 6, 8r et 9) dans la liste des variables, en les sélectionnant, toujours avec la flèche, mais dans le sens inverse.

Les options choisies restent cochées (extraction, demande du KMO, etc.) et il n'est donc pas nécessaire de recommencer cette procédure. Cliquez sur **OK**.

Les résultats de cette deuxième AFE apparaissent, toujours dans l'onglet résultats, à la figure 3.10.

La solution est maintenant, comme dans la théorie, unidimensionnelle, mais elle ne parvient toujours pas à expliquer plus de 50 % de la variance. Dès lors, l'item 3r dont la communalité est insuffisante (0,38) est supprimé.

Nous rappelons donc la boîte de dialogue (voir figure 3.11) et nous faisons passer l'item ambition3r dans la liste des variables. Puis nous cliquons sur **OK.**

Figure 3.9

Demande d'analyse factorielle (bis).

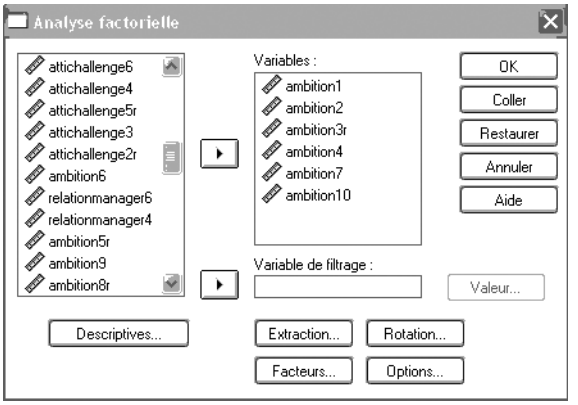


Figure 3.10

Interprétation des résultats de l'AFE : KMO, test de Bartlett et communalités (bis).

Qualité de représentation

	Initial	Extraction
ambition1	1,000	,506
ambition2	1,000	,470
ambition3r	1,000	,389
ambition4	1,000	,511
ambition7	1,000	,547
ambition10	1,000	,575

Méthode d'extraction : Analyse en composantes principales.

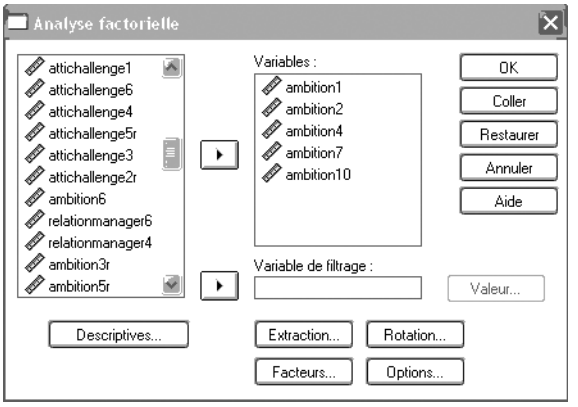
Variance totale expliquée

	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
Composante	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	2,999	49,991	49,991	2,999	49,991	49,991
2	,747	12,451	62,441			
3	,638	10,635	73,076			
4	,616	10,274	83,350			
5	,522	8,706	92,056			
6	,477	7,944	100,000			

Méthode d'extraction : Analyse en composantes principales.

Figure 3.11

Demande d'AFE n° 3.



Les résultats de cette troisième AFE apparaissent à la suite des autres, dans l'onglet résultats, à la figure 3.12.

Figure 3.12

Interprétation des résultats de l'AFE : KMO, test de Bartlett et communalités n° 3.

Indice KMO et test de Bartlett

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,826
Test de sphéricité de Bartlett	Khi-deux approximé	868,722
	ddl	10
	Signification de Bartlett	,000

Qualité de représentation

	Initial	Extraction
ambition1	1,000	,514
ambition2	1,000	,457
ambition4	1,000	,527
ambition7	1,000	,582
ambition10	1,000	,608

Méthode d'extraction : Analyse en composantes principales.

Variance totale expliquée

	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
Composante	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	2,688	53,752	53,752	2,688	53,752	53,752
2	,674	13,475	67,226			
3	,623	12,468	79,695			
4	,538	10,758	90,453			
5	,477	9,547	100,000			

Méthode d'extraction : Analyse en composantes principales.

Cette dernière solution unidimensionnelle permet d'expliquer 53,72 % de la variance ; les communalités et les composantes de chaque item sont respectivement supérieures à 0,45 et 0,67. La matrice des composantes (voir figure 3.13) indique que tous les items ont un coefficient structurel ou > à 0,65.

Les items 1, 2, 4, 7 et 10 sont donc conservés pour la suite de l'analyse.

Figure 3.13

Interprétation des résultats de l'AFE : matrice des composantes (coefficients structurels) n° 3.

Matrice des composantes[a]

	Composante
	1
ambition1	,717
ambition2	,676
ambition4	,726
ambition7	,763
ambition10	,780

Méthode d'extraction : Analyse en composantes principales.

a. 1 composantes extraites

À ce stade nous procédons à l'examen de la fiabilité de l'échelle avec le coefficient alpha de Cronbach.

Le calcul du coefficient alpha de Cronbach pour vérifier la fiabilité d'une échelle

Nous continuons le processus d'épuration des données avec le même exemple d'échelle de mesure de l'ambition du vendeur. Nous avons vu que le coefficient alpha de Cronbach était un indicateur de la cohérence interne d'une échelle de mesure.

Allez dans le menu **Analyse > Positionnement > Analyse de fiabilité.** La boîte de dialogue de la figure 3.14 apparaît.

Transférez les items sélectionnés destinés à mesurer l'ambition à l'aide de l'analyse factorielle exploratoire en les sélectionnant chacun à leur tour puis en cliquant sur la flèche.

Avant de lancer le calcul de l'alpha de Cronbach, cliquez sur le bouton **Statistiques,** la boîte de dialogue de la figure 3.15 apparaît alors. Nous demandons l'alpha pour chaque item, pour l'échelle et l'échelle sans l'item.

Figure 3.14

Test de la fiabilité de cohérence interne avec le coefficient alpha de Cronbach.

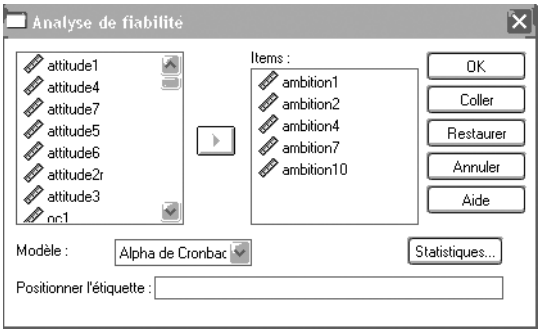
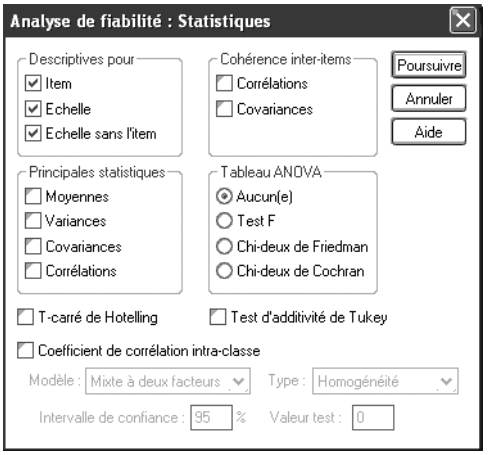


Figure 3.15

Choix des statistiques pour le calcul du coefficient alpha de Cronbach.



Les résultats apparaissent dans l'onglet résultats, à la figure 3.16.
Le coeficient alpha de Cronbach apparaît dans le deuxième tableau.

Figure 3.16

Résultats du calcul du coefficient alpha de Cronbach.

Récapitulatif de traitement des observations

		N	%
Observations	Valide	705	94,4
	Exclus[a]	42	5,6
	Total	747	100,0

a. Suppression par liste basée sur toutes les variables de la procédure.

Statistiques de fiabilité

Alpha de Cronbach	Nombre d'éléments
,782	5

Statistiques d'item

	Moyenne	Ecart-type	N
ambition1	2,7716	1,19554	705
ambition2	3,0241	1,10820	705
ambition4	3,0043	1,28308	705
ambition7	2,4695	1,15404	705
ambition10	2,7348	1,06160	705

Le premier tableau présente la moyenne et la variance de l'échelle en cas de suppression de calcul des items (colonnes 1 et 2), la corrélation de chaque item aux autres (colonne 3) et l'alpha de Cronbach en cas de supression d'un item.

Figure 3.17

Résultat du calcul du coefficient alpha de Cronbach.

Statistiques de total des éléments

	Moyenne de l'échelle en cas de suppression d'un élément	Variance de l'échelle en cas de suppression d'un élément	Corrélation complète des éléments corrigés	Alpha de Cronbach en cas de suppression de l'élément
ambition1	11,2326	12,142	,538	,748
ambition2	10,9801	12,872	,497	,760
ambition4	11,0000	11,588	,552	,745
ambition7	11,5348	11,982	,593	,729
ambition10	11,2695	12,345	,614	,725

Statistiques d'échelle

Moyenne	Variance	Ecart-type	Nombre d'éléments
14,0043	18,055	4,24916	5

Cette échelle présente une fiabilité de cohérence interne acceptable (alpha = 0,78). Il n'est pas possible d'améliorer l'alpha en éliminant un ou plusieurs items (*cf.* colonne droite du tableau alpha en cas de suppression de l'élément). Nous gardons donc les cinq items (1, 2, 4, 7 et 10) pour mesurer l'ambition.

Pour aller plus loin

Evrard Y., Pras B., et Roux E., *Market. Études et recherches en marketing*, Nathan, Paris, 2003.

Gerbing D. W., Anderson J. C., « An updated paradigm for scale development incorporating unidimensionality and its assessment », *Journal of Marketing Research*, 25, 1988, p. 186-192.

Hair J. F., Anderson R. E., Tatham R. L., Black W. C., *Multivariate Data Analysis*, 4e éd., Prentice Hall International, New Jersey, 2006.

Exercices

Exercice 1 Analyse d'une AFC

Énoncé

Il existe différentes sources de satisfaction au travail, ces valences sont évaluées en posant la question : « Quelle importance accordez-vous à ces aspects de votre travail ? » (collecte des réponses à l'aide d'une échelle de Likert en cinq points allant de « Pas du tout » à « Très important »).

VAL1 - Une augmentation de votre sentiment réussite

VAL2 - Le sentiment que vous utilisez bien vos compétences

VAL3 - Votre satisfaction personnelle

VAL4 - L'occasion de développer des relations avec les autres employés de l'entreprise

VAL5 - De meilleures relations de travail avec votre manager

VAL6 - De meilleures relations avec les autres commerciaux

VAL7 - L'implication dans la formation des autres employés

VAL8 - Plus d'autonomie de la part de votre manager

VAL9 - Davantage de respect de la part de vos collègues

VAL10 - Une baisse des réclamations de la part de vos clients

VAL11 - La reconnaissance de vos clients sur le fait que vous les avez bien conseillés

VAL12 - Plus d'opportunités de développer des contacts clients

VAL13 - De meilleures relations avec vos clients

VAL14 - Une augmentation de vos revenus

VAL15 - Plus d'influence sur les décisions de votre manager

VAL16 - Recevoir la reconnaissance de votre hiérarchie

VAL17 - Une augmentation de votre prestige personnel

VAL18 - La chance d'être muté dans une agence ayant plus de potentiel

VAL19 - L'évolution vers un poste de management

Nous collectons aussi la valence par rapport à la victoire au challenge :

VALVI1 - La victoire à ce challenge

VALVI2 - Être parmi les gagnants du challenge en question

Suite à une première AFE, seuls les items apparaissant en gras ont été conservés.

Question : décrivez les résultats de l'AFC (voir figures 3.18, 3.19, 3.20 et 3.21).

Figure 3.18
Résultats de l'AFC (1).

Indice KMO et test de Bartlett

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,867
Test de sphéricité de Bartlett	Khi-deux approximé	4276,977
	ddl	78
	Signification de Bartlett	,000

Qualité de représentation

	Initial	Extraction
val2	1,000	,543
val4	1,000	,611
val6	1,000	,622
val9	1,000	,435
val10	1,000	,585
val11	1,000	,629
val12	1,000	,460
val13	1,000	,635
val14	1,000	,440
val15	1,000	,401
val18	1,000	,461
valvi1	1,000	,783
valvi2	1,000	,759

Méthode d'extraction : Analyse en composantes principales.

Figure 3.19
Résultats de l'AFC (2).

Variance totale expliquée

	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus			Somme des carrés des facteurs retenus pour la rotation		
Composante	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	5,256	40,432	40,432	5,256	40,432	40,432	4,287	32,974	32,974
2	2,108	16,216	56,647	2,108	16,216	56,647	3,078	23,674	56,647
3	,858	6,602	63,250						
4	,798	6,137	69,386						
5	,659	5,066	74,452						
6	,611	4,700	79,152						
7	,600	4,619	83,771						
8	,525	4,035	87,807						
9	,513	3,947	91,753						
10	,357	2,747	94,501						
11	,331	2,544	97,044						
12	,205	1,579	98,623						
13	,179	1,377	100,000						

Méthode d'extraction : Analyse en composantes principales.

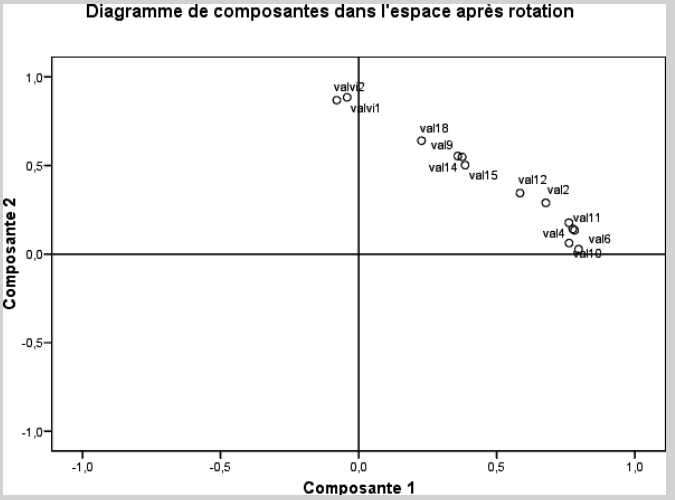
Figure 3.20
Résultats de l'AFC (3).

Matrice des composantes après rotation[a]

	Composante	
	1	2
val2	,678	,289
val4	,761	,177
val6	,776	,141
val9	,359	,553
val10	,762	,062
val11	,782	,134
val12	,584	,344
val13	,796	,027
val14	,375	,548
val15	,386	,502
val18	,228	,640
valvi1	-,042	,884
valvi2	-,079	,868

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation de Kaiser.
a. La rotation a convergé en 3 itérations.

Figure 3.21
Résultats de l'AFC (4).



Solution

Les résultats de l'AFC sont comparables à ceux qui peuvent être obtenus à l'aide d'une ACP. Le premier facteur représente 32,97 % de la variance, le second compte pour 23,67 %.

Sur l'axe horizontal, nous trouvons les relations que le vendeur entretient avec ses clients, ses pairs. Sur l'axe vertical (deuxième composante), nous trouvons la valence pour la victoire au challenge. Nous observons que les items 18 et 9, qui portent sur les promotions, sont proches de cet axe. Le challenge serait donc associé aux opportunités de carrière, au respect des pairs. Sur cet axe, nous retrouvons les aspects relationnels du travail, en interne (avec les collègues) ou en externe (avec les clients).

Exercice 2 Analyse de la validité et de la fiabilité

Énoncé

Il n'existe pas d'échelle dans la littérature pour mesurer l'attitude générale à l'égard des challenges de vente. L'échelle de mesure de cette attitude a été créée grâce à une étude qualitative : 7 items ont été créés puis utilisés dans un questionnaire. Les réponses sont recueillies auprès de 747 commerciaux sur une échelle de Likert à cinq grades.

Une analyse factorielle exploratoire (ACP), puis un test de la fiabilité de cette échelle sont réalisés (voir figures 3.22, 3.23 et 3.24).

À partir de ces analyses, répondez aux questions suivantes :

1. Cette échelle est-elle multidimensionnelle ?
2. Faut-il conserver tous les items de cette échelle ?
   Sinon quel(s) item(s) éliminieriez-vous ? Comment prenez-vous cette décision ?
3. Cette échelle de mesure est-elle fiable ?

Figure 3.22
Résultats de l'AFE (1).

Indice KMO et test de Bartlett

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,888
Test de sphéricité de Bartlett	Khi-deux approximé	1234,588
	ddl	21
	Signification de Bartlett	,000

Qualité de représentation

	Initial	Extraction
atig1	1,000	,727
atig2r	1,000	,567
atig3	1,000	,738
atig4	1,000	,456
atig5	1,000	,236
atig6	1,000	,515
atig7	1,000	,483

Méthode d'extraction : Analyse en composantes principales.

Figure 3.23
Résultats de l'AFE (2).

Variance totale expliquée

	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
Composante	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	3,722	53,174	53,174	3,722	53,174	53,174
2	,859	12,274	65,448			
3	,684	9,772	75,220			
4	,585	8,355	83,575			
5	,503	7,186	90,760			
6	,355	5,076	95,836			
7	,291	4,164	100,000			

Méthode d'extraction : Analyse en composantes principales.

Matrice des composantes[a]

	Composante
	1
atig1	,852
atig2r	,753
atig3	,859
atig4	,676
atig5	,486
atig6	,717
atig7	,695

Méthode d'extraction : Analyse en composantes principales.
a. 1 composantes extraites

Figure 3.24
Résultats de l'AFE (3).

Fichier Edition Affichage Données Transformer Insérer Format Analyse Graphes Outils Fenêtre Aide

Statistiques de fiabilité

Alpha de Cronbach	Nombre d'éléments
,842	7

Statistiques d'item

	Moyenne	Ecart-type	N
atig1	3,0283	1,16469	460
atig2r	2,9565	1,28264	460
atig3	3,1087	1,24071	460
atig4	2,7522	1,30505	460
atig5	2,8109	1,40070	460
atig6	2,8000	1,32481	460
atig7	2,6978	1,21626	460

Statistiques de total des éléments

	Moyenne de l'échelle en cas de suppression d'un élément	Variance de l'échelle en cas de suppression d'un élément	Corrélation complète des éléments corrigés	Alpha de Cronbach en cas de suppression de l'élément
atig1	17,1261	30,093	,754	,798
atig2r	17,1978	30,599	,623	,816
atig3	17,0457	29,299	,763	,794
atig4	17,4022	31,282	,555	,826
atig5	17,3435	33,050	,377	,856
atig6	17,3543	30,765	,583	,822
atig7	17,4565	31,778	,571	,824

Solution

1. L'analyse en composantes principales atteste de la nature unidimensionnelle de ce concept. Cette solution factorielle parvient à expliquer plus de 53 % de la variance totale. Le test KMO, tout à fait satisfaisant (0,88) valide cette solution factorielle.

2. Non, il ne faut pas conserver tous les items.

L'examen des communalités des énoncés indique que l'item atig5, dont l'indice de communalité (0,23) et le poids factoriel (0,48) sont faibles, affaiblit la validité de cette échelle. Cet item doit par conséquent, être éliminé pour la suite des analyses.

3. Cette échelle est fiable puisque le coefficient alpha de Cronbach dépasse 0,8 (0,84).

Toutefois la fiabilité peut être meilleure si l'item atig5 est éliminé (l'alpha monte à 0,85).

2.1 Exercice 3 : Générer une carte perceptuelle par l'AFC

Énoncé

Une enquête portant sur les perceptions de différentes marques de voitures a été réalisée auprès de consommateurs. Les individus ont évalué 10 marques d'après 15 critères, notés sur des échelles de Likert de 1 à 9. Les variables perceptuelles sont les suivantes :

Notoriété	Ergonomie
Finition	Prestige
Qualité	Familial
Confort	Économique
Nouveauté	Image
Qualité-prix	Innovation
Robustesse	Sportif
Spacieux	

Les résultats de l'enquête, c'est-à-dire la moyenne des scores obtenus à chaque variable, sont représentés dans le fichier « Facto.sav ». Sur ces données, une analyse factorielle exploratoire peut permettre d'identifier les perceptions de consommateurs, mais aussi de représenter les marques en fonction de ces perceptions dans ce que l'on nomme une carte perceptuelle, ou mapping perceptuel.

1. Générez l'analyse factorielle sur les données de l'étude.
2. Interprétez l'analyse factorielle. Quelles conclusions tirez-vous de cette analyse ?

Solution

1. Pour commander l'analyse factorielle, sélectionnez le menu **Analyse** > **Factorisation** > **Analyse factorielle** et faites passer les variables à factoriser dans la partie **Variables** avec la flèche (voir figure 3.25).

Ensuite, dans l'onglet **Descriptives** (voir figure 3.26), la case **Structure initiale** est déjà cochée (elle donne les communautés, valeurs propres et pourcentage de variance expliqués par chaque dimension). Dans la partie **Matrice des corrélations**, cochez les cases **Coefficients** et **Reconstituée**.

Figure 3.25

Commande de l'analyse factorielle.

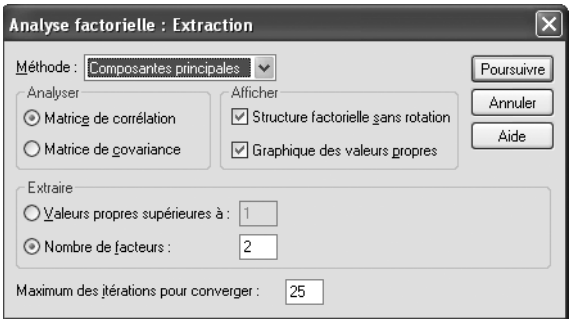
Figure 3.26

Options d'Analyse factorielle.

Dans l'onglet **Extraction** (voir figure 3.27), la case **Matrice de corrélation** est cochée et la méthode en **Composantes principales** sélectionnée. Cochez **Graphique des valeurs propres** et **Structure factorielle sans rotation**. Sélectionnez un **nombre de facteurs** égal à 2, afin de générer une carte à deux dimensions.

Figure 3.27

Méthode d'extraction de l'analyse factorielle.



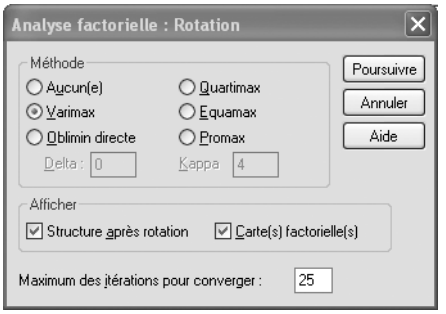
Dans l'onglet **Rotation**, choisissez **Varimax** et cliquez sur **Carte factorielle** comme l'indique la figure 3.28.

Une fois tous les paramètres définis, cliquez sur **OK** pour lancer l'analyse factorielle.

2. L'interprétation de l'analyse factorielle s'établit à l'aide des tableaux qui apparaissent dans la partie Résultats :

Figure 3.28

Choix de la méthode de rotation de l'analyse factorielle.



Le tableau de la variance totale expliquée (voir figure 3.29) présente les deux dimensions qui résument l'information. La première dimension permet d'expliquer 46,32 % de la variance du phénomène, c'est-à-dire que les variables qui composent cette première dimension synthétisent 51,63 % du phénomène. La seconde dimension explique 23,95 % de la variance. Les deux dimensions expliquent plus de 70 % de la variance totale. On conseille en général d'arrêter l'extraction de facteurs lorsque 60 % de variance cumulée a été extraite (Hair *et al.*, 1998). Cette variance cumulée indique que la réduction des variables à deux composantes permet de conserver l'essentiel du phénomène mesuré par les quinze variables perceptuelles initiales. Notre représentation du phénomène est donc de qualité.

Figure 3.29

Les résultats de l'analyse factorielle : la variance totale expliquée.

Variance totale expliquée

	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus			Somme des carrés des facteurs retenus pour la rotation		
Composante	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	7,745	51,634	51,634	7,745	51,634	51,634	6,948	46,323	46,323
2	2,795	18,635	70,270	2,795	18,635	70,270	3,592	23,946	70,270
3	2,062	13,750	84,019						
4	1,276	8,510	92,529						
5	,442	2,948	95,477						
6	,388	2,585	98,062						
7	,201	1,343	99,405						
8	,069	,459	99,864						
9	,020	,136	100,000						
10	4,35E-016	2,90E-015	100,000						
11	1,50E-016	1,00E-015	100,000						
12	6,23E-017	4,15E-016	100,000						
13	-1,3E-016	-8,6E-016	100,000						
14	-2,3E-016	-1,5E-015	100,000						
15	-9,4E-016	-6,2E-015	100,000						

Méthode d'extraction : Analyse en composantes principales.

La qualité de la représentation (voir figure 3.30) permet de vérifier si les variables initiales sont bien prises en compte par les variables extraites. Ici, la qualité de représentation ou communalité de la variable « notoriété » est de 0,989. Ce qui signifie que 98,9 % de la variance de la variable est prise en compte par l'une des deux dimensions extraites. Dans cet exemple, les variables « nouveauté » et « innovation » ne sont pas bien représentées.

La matrice des composantes (voir figure 3.31) montre les dimensions extraites (deux dimensions) avec les composantes. Chaque colonne correspond à une dimension extraite contenant les coefficients ou composantes qui peuvent s'interpréter comme des coefficients de corrélation.

La qualité et le confort sont ainsi reliés à la dimension 1, alors que le rapport qualité-prix ou la dimension économique du modèle sont reliés à la dimension 2. On passe donc en revue les coefficients afin d'identifier les variables reliées à chacune des dimensions. De cette manière, la matrice des composantes permet de nommer les dimensions extraites par l'étude des composantes. La première composante relève de l'image perçue (prestige à gauche de l'axe, et familial à droite de l'axe ; voir figure 3.32) ; la seconde composante relève du rapport qualité-prix perçu.

Figure 3.30

Les résultats de l'analyse factorielle : la qualité de représentation.

Qualité de représentation

	Initial	Extraction
Notoriété	1,000	,989
Finition	1,000	,761
Qualité	1,000	,878
Confort	1,000	,916
Ergonomie	1,000	,560
Qualité-Prix	1,000	,844
Robustesse	1,000	,854
Sportif	1,000	,878
Economique	1,000	,785
Prestige	1,000	,864
Familial	1,000	,578
Nouveauté	1,000	,029
Image	1,000	,886
Innovation	1,000	,236
Spacieux	1,000	,484

Méthode d'extraction : Analyse en composantes principales.

Figure 3.31

Les résultats de l'analyse factorielle : la matrice des composantes.

Matrice des composantes après rotation[a]

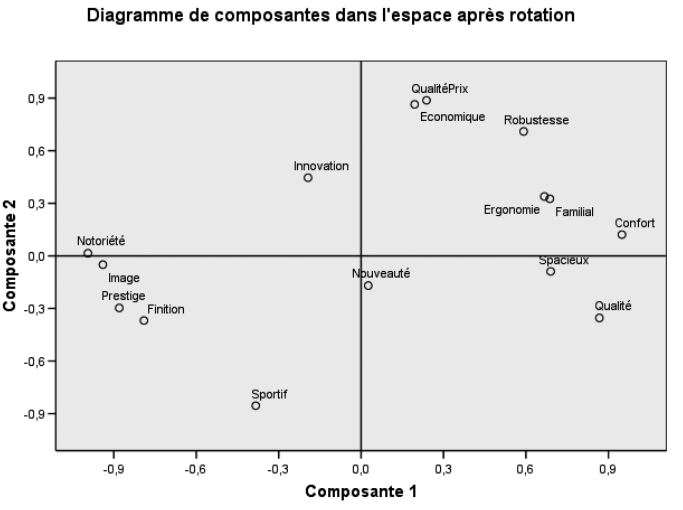
	Composante	
	1	2
Notoriété	-,994	,014
Finition	-,791	-,369
Qualité	,867	-,354
Confort	,949	,121
Ergonomie	,667	,338
Qualité-Prix	,238	,887
Robustesse	,592	,710
Sportif	-,384	-,855
Economique	,195	,864
Prestige	-,880	-,297
Familial	,687	,326
Nouveauté	,026	-,170
Image	-,940	-,050
Innovation	-,193	,446
Spacieux	,690	-,089

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation de Kaiser.
a. La rotation a convergé en 3 itérations.

Le diagramme des composantes correspond à la représentation graphique de la matrice des composantes (voir figure 3.32).

Figure 3.32

Les résultats de l'analyse factorielle : le diagramme des composantes.

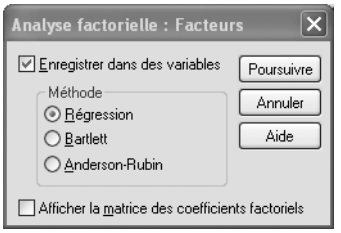


La matrice des composantes permet également de calculer les coordonnées pour représenter graphiquement les individus par rapport aux dimensions extraites. On peut ainsi comparer la position de chacune des observations, en d'autres termes, positionner les marques dans l'espace factoriel créé.

Afin de retrouver les marques sur chacun de ces axes, vous devez relancer l'analyse factorielle : **Analyse** > **Analyse factorielle** puis, dans l'onglet **Facteurs**, cocher **Enregistrer dans des variables** et la méthode **Régression** (voir figure 3.33).

Figure 3.33

La commande Analyse factorielle : représentation graphique d'individus.



Deux nouvelles variables sont alors créées dans l'éditeur de données (voir figure 3.34). Elles donnent les scores factoriels, c'est-à-dire pour chaque individu (chaque marque) sa moyenne sur chacune des deux dimensions. On peut constater par exemple que la Citroën C4 est reliée à la dimension Image. C'est ce que nous allons maintenant voir à l'aide d'un graphique.

Figure 3.34

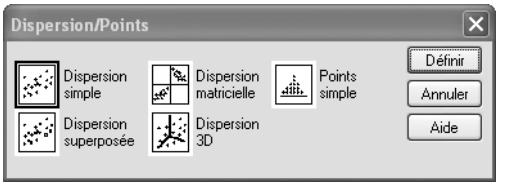
La représentation graphique d'individus dans l'analyse factorielle.

	Modèle	FAC1_1	FAC2_1
1	Série 1 (BMW)	-,84125	-,01247
2	147 (AlfaRomeo)	,26090	-1,74462
3	Focus (Ford)	-1,00574	1,66422
4	Megane (Renault)	1,29757	,04380
5	Golf (Volkswagen)	-,53244	,37056
6	Classe A (Mercedes)	-,52802	-1,11059
7	C4 (Citroën)	1,24942	,78323
8	A3 Sportback (Audi)	-1,04583	-,31361
9	307 (Peugeot)	1,14540	,31947

Pour commander le graphique, sélectionnez le menu **Graphes** > **Boîtes de dialogue héritées** > **Dispersion/Points**, puis cliquez sur **Définir** (voir figure 3.35).

Figure 3.35

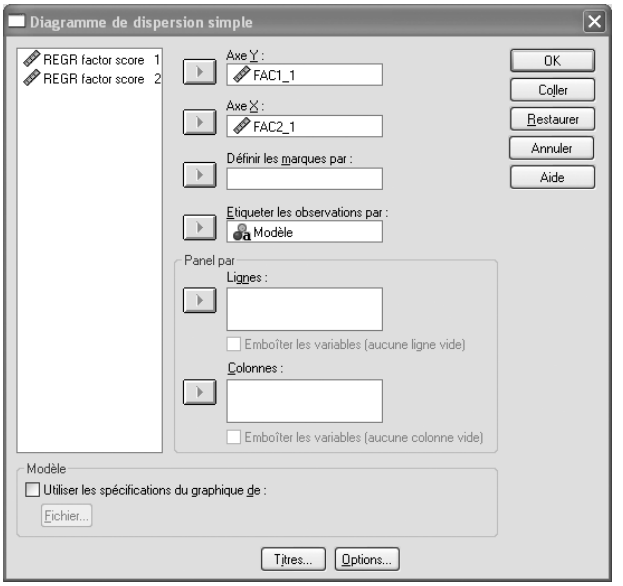
Commande d'une représentation graphique d'individus dans l'analyse factorielle.



Ensuite faites passer les facteurs créés dans les axes Y et X et, afin d'afficher chaque marque, faites glisser la variable « modèle » vers **Etiqueter les observations par** (voir figure 3.36).

Figure 3.36

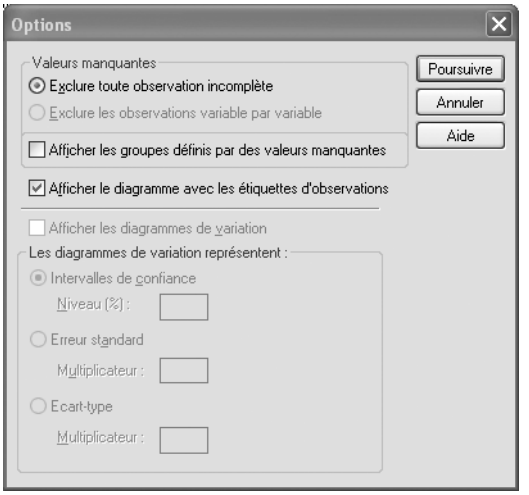
Commande d'une représentation graphique d'individus dans l'analyse factorielle (suite).



Vous devez également cliquer sur l'onglet **Options** et cocher **Afficher le diagramme avec les étiquettes d'observations** pour les faire apparaître (voir figure 3.37).

Figure 3.37

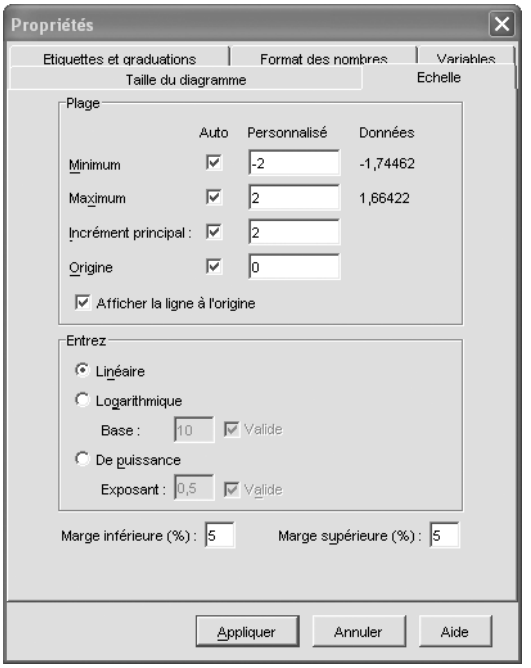
Commande d'une représentation graphique d'individus dans l'analyse factorielle (suite).



Ensuite, cliquez sur **OK** : le graphique n'est cependant pas très lisible car les axes n'apparaissent pas. Pour y remédier, double-cliquez sur le graphique pour ouvrir l'éditeur de diagramme (voir chapitre 8), activez le diagramme en cliquant une nouvelle fois dessus et sélectionnez dans le menu **Édition** la ligne de référence de l'axe X ; la position de l'axe 0 est présélectionnée. Dans le menu **Propriétés**, sélectionnez **Afficher la ligne à l'origine**, dans l'onglet **Echelle** (voir figure 3.38). Recommencez cette opération pour l'axe Y.

Figure 3.38

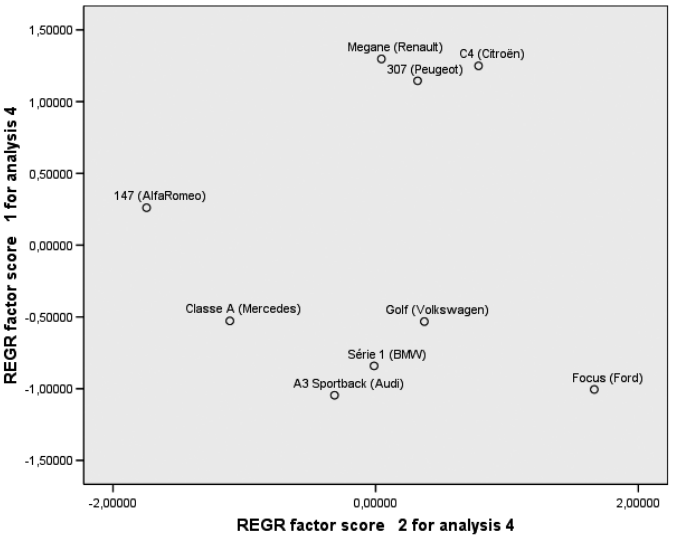
Commande d'une représentation graphique d'individus dans l'analyse factorielle (suite).



Le graphique suivant (voir figure 3.39) apparaît alors, représentant les modèles de véhicules en fonction des perceptions déclarées des consommateurs interrogés. Ce type de représentation peut servir à positionner les offres concurrentes sur le marché.

Figure 3.39

Représentation graphique d'individus dans l'analyse factorielle.



Chapitre 4

Segmenter

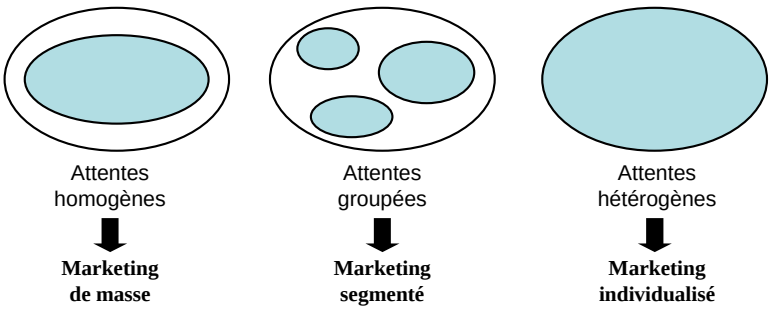
De nombreuses enquêtes en marketing ont pour objet de classer des individus en groupes homogènes, afin, par exemple, de procéder à une segmentation du marché. Comme l'analyse factorielle (voir chapitre 3), l'analyse typologique permet de réduire le nombre des observations en les regroupant en des classes (ou types) homogènes et différenciées. (Evrard *et al.*, 2003). Cependant, contrairement à l'analyse factorielle, les résultats peuvent fortement diverger en fonction des choix effectués. Nous verrons, dans ce chapitre, les concepts clés liés à cette méthode, les mesures statistiques associées, ainsi que les principaux éléments de sa mise en œuvre.

1 Fondements

La réalisation d'une typologie, ou encore d'une taxinomie, a été pendant longtemps le principe fondateur de la science moderne. Il s'agissait alors de décrire le monde afin de le comprendre. En français, les termes pour décrire ce principe de classification des individus – typologie, taxinomie, segmentation, classification, catégorisation –, sont relativement ambigus. En anglais, en revanche, le terme *clustering* rend compte à la fois du fait de classer, c'est-à-dire de faire émerger des groupes d'individus, mais également du principe de classification, c'est-à-dire de l'affectation des individus aux différents groupes. L'analyse typologique, terme générique que nous retiendrons dans ce chapitre, est au cœur de la démarche marketing. Elle peut être utilisée pour simplifier la lecture des données en regroupant des observations ayant des caractéristiques communes, ou encore pour faire émerger des groupes d'individus homogènes des données collectées. Cette approche est fréquemment retenue en marketing, où l'une des premières décisions stratégiques consiste à agréger des segments de marché en fonction des attentes des consommateurs afin de définir les choix de marchés possibles. Le marketing considère, en effet, que les marchés sur lesquels il opère peuvent être constitués d'attentes plus ou moins homogènes, qu'il s'agira de rendre intelligibles. On parle de **marketing de masse** lorsque les attentes sont homogènes, de **marketing individualisé** lorsque les attentes sont fortement hétérogènes, et de **marketing segmenté** lorsque les attentes sont groupées (voir figure 4.1).

Figure 4.1

Les attentes en marketing.



Le rôle du chargé d'étude dans cette perspective est de créer *ex nihilo* des groupes de consommateurs similaires entre eux mais différents des autres segments. Le principe de la segmentation, c'est-à-dire « *le fait de former des groupes de clients homogènes* », est directement fondé sur des caractéristiques propres des individus (les consommateurs en l'occurrence) qui nous indiquent pourquoi les segments diffèrent. Ces critères vont permettre au chargé d'étude d'identifier et de rapprocher les membres d'un segment. Pour qu'ils soient utiles, ces critères de segmentation doivent permettre de générer des segments distincts, en d'autres termes qui ne se recoupent pas. Il n'existe pas de segmentation optimale : la segmentation est un moyen par lequel on pourra, par exemple, identifier les cibles d'une campagne de mailing, orienter une extension de gamme de produits, définir le message publicitaire adapté à un profil de consommateurs, etc. Il existe donc plusieurs résultats possibles lors d'une segmentation, et le chargé d'étude devra réaliser des choix, afin de concilier exigence statistique et besoin d'opérationalité de la décision.

Pour regrouper des consommateurs, on considérera des variables de segmentation :

- **géographiques** : région, type d'habitat, type d'agglomération, etc. ;
- **sociodémographiques** : âge, sexe, taille du foyer, revenus, catégorie socioprofessionnelle, niveau d'éducation, etc. ;
- **psychographiques** : style de vie, rapport au temps, personnalité, valeurs, etc. ;
- **comportementales** : attitudes, préférences, comportement d'achat (Récence, Fréquence, Montant), etc.

Dans un cadre de marketing industriel ou *B to B* (pour segmenter des entreprises), on pourra utiliser l'activité exercée par l'entreprise (le code NAF de l'Insee par exemple), la taille de l'entreprise (nombre d'employés, chiffre d'affaires), la localisation, la structure (divisions, magasins propres/franchises), etc.

EXEMPLE

GDF et la segmentation à 360° [1]

Pour faire face à l'ouverture du marché de l'énergie mise en place en France le 1er juillet 2007, GDF mène depuis quelques années une réflexion approfondie sur ses méthodes de segmentation, afin de mieux connaître ses clients et leurs comportements, et surtout limiter leur départ vers la concurrence. Cette segmentation repose sur une base de données recoupant les informations issues de données commerciales, marketing (CRM) et d'administration des ventes (niveau de consommation, facturation et paiement). En défragmentant les données clients, GDF a affiné sa vision de la valeur économique de son portefeuille clients. La conséquence est une nouvelle approche de la segmentation clients : quinze segments de clientèle ont été constitués et agrégés en cinq macrosegments : les clients à convaincre, à conforter, à consolider, à observer et à tolérer. Une stratégie différenciée a ensuite été mise en place pour chacun des quinze segments (fidélisation, promotions, etc.).

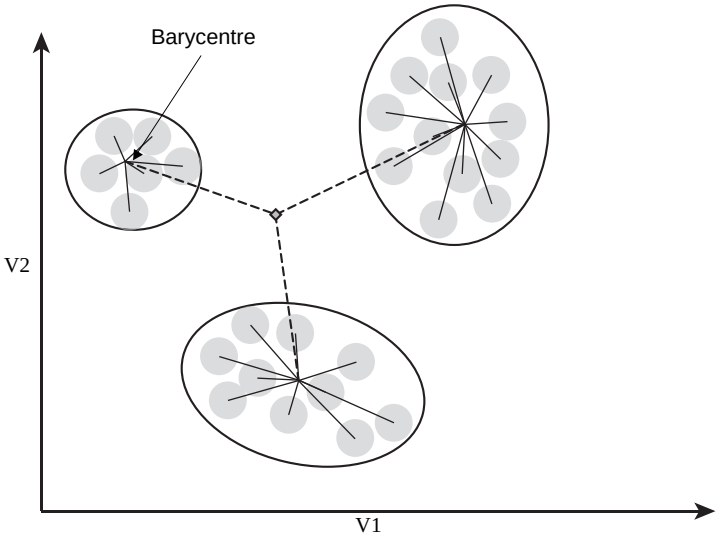
Nous nous intéresserons ici aux principales approches, les plus diffusées dans la pratique et les plus aisées à mettre en œuvre en termes d'analyse de données, que nous regrouperons sous le terme d' «analyse typologique ». L'analyse typologique peut être définie de la façon suivante : « Étant donné un ensemble d'objets (ou d'individus) décrits par un certain nombre de caractéristiques (ou variables), constituer des groupes (ou types) d'objets tels que les objets soient les plus similaires possibles au sein d'un groupe et que les groupes soient aussi dissemblables que possible ; la ressemblance ou la dissemblance étant mesurée sur l'ensemble des variables décrivant les objets » (Evrard *et al.*, 2003). Ces regroupements sont effectués en fonction de variables, dont on peut distinguer deux types : des variables comportementales pour classer les individus dans les segments, et des variables d'identification afin de pouvoir interpréter les groupes ainsi générés. Le choix des variables employées est hautement important. En effet, lorsque l'on procédera à l'analyse typologique, les résultats devront montrer une forte homogénéité intragroupe (proximité des mesures au sein d'un groupe), et une forte hétérogénéité intergroupe (distance entre les groupes) comme le montre la figure 4.2.

La figure 4.2. représente un nuage de points issu du croisement de deux variables, soit par exemple *l'âge (V1)* et *la fréquence d'achat (V2)* d'un produit X. Le centre de gravité du nuage de points est représenté par le point au centre des trois segments. Comme on peut le constater, trois grands groupes émergent lorsque l'on croise ces deux variables. Chaque point représentant la combinaison des deux variables pour une observation, on remarque

1. Adapté de « GDF : Fidéliser et conquérir de nouveau marchés grâce au data mining », *Decisio*, 43, juin 2007.

Figure 4.2

La segmentation.



que les individus regroupés sont relativement homogènes, proches, et différents – c'est-à-dire distants – des autres membres des deux autres segments en termes d'âge et de fréquence d'achat. Le chargé d'étude pourra aisément recommander des stratégies pour servir ces trois segments en termes de promotion ou de message publicitaire, par exemple. On peut voir, en effet, que les consommateurs les plus âgés, puis les plus jeunes, achètent le plus fréquemment, alors que les consommateurs des classes d'âge intermédiaires achètent le moins fréquemment. Malheureusement, il est extrêmement rare, pour ne pas dire impossible, que de tels segments émergent dans la pratique ; l'objectif de la segmentation sera donc de créer des groupes distincts les uns des autres, mais dont les caractéristiques seront proches au sein même des groupes. En d'autres termes, il s'agira de diminuer le plus possible les traits pleins sur le schéma, à l'intérieur des groupes, et d'augmenter au maximum les traits en pointillés afin de bien distinguer les segments les uns des autres.

Le chargé d'étude, afin d'éviter les erreurs liées à ce type d'exercice, devra, au-delà d'une maîtrise des principales mesures associées à la typologie, être en mesure de décider du nombre de segments satisfaisant les objectifs de l'enquête, et interpréter correctement le contenu de chacun des segments générés. Le problème que pose l'analyse typologique, que nous avons définie comme une technique d'analyse de données multivariée utilisée pour segmenter des populations, est précisément qu'elle repose sur un classement des individus et non une mesure des variables comme c'est le cas pour les autres tests statistiques. En d'autres termes, il n'existe pas une mais des analyses typologiques. De nombreuses possibilités sont donc offertes au chargé d'étude. Cet outil présente une grande flexibilité d'utilisation, mais également une importante complexité car le risque d'obtenir des résultats pertinents mais influencés par les procédures de calcul retenues et non par les données est important. C'est ce que l'on nomme le **risque d'artefact**.

2 Concepts associés

Il existe un certain nombre de concepts associés à l'analyse typologique. Deux dimensions principales doivent être abordées : les mesures statistiques de distance entre les individus et le processus de constitution des groupes qui sera sélectionné par l'analyste.

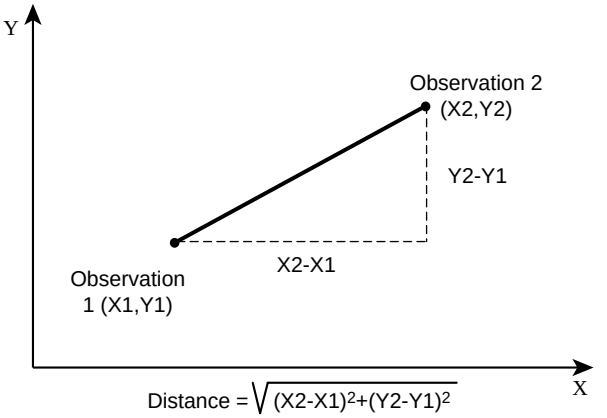
2.1 Mesures de distance

Le concept de distance est aisément compréhensible si l'on se réfère à la représentation des données sous forme de points dans un espace tel que représenté par la figure 4.2. On peut faire un parallèle ici avec l'analyse factorielle que nous avons étudiée au chapitre 3. Lors d'une analyse factorielle, la matrice des corrélations est employée pour regrouper des variables deux à deux au sein de différents facteurs. La corrélation représente le lien entre deux variables parmi toutes les observations. L'analyse factorielle regroupe donc au sein d'un même facteur toutes les variables ayant de fortes corrélations entre elles. La démarche est un peu similaire lors d'une analyse typologique. La mesure de distance est calculée pour chaque paire d'objets sur la base de leurs caractéristiques telles que spécifiées par l'analyste. Ainsi, chaque objet peut être comparé par le biais de cette mesure de distance afin de former des groupes homogènes.

Les mesures de distance, comme leur nom l'indique, représentent l'éloignement entre deux observations en fonction de leurs caractéristiques, une valeur élevée représentant une faible proximité. Cette distance peut être convertie en mesure de proximité en inversant la relation. La principale mesure utilisée est la **distance euclidienne**, qui consiste à calculer la racine carrée de la somme des carrés des différences entre les valeurs de chaque variable. La figure 4.3 illustre cette mesure.

Figure 4.3

Illustration de la distance euclidienne [1].



1. Adapté de Hair *et al.*, 2006, p. 575.

On peut voir sur la figure 4.3 que la distance euclidienne mesure la distance (ou la proximité) entre deux observations dont les coordonnées dans l'espace sont données par les valeurs des variables X et Y. Pour l'observation 1, ces coordonnées sont (X1, Y1) et (X2, Y2) pour l'observation 2. La distance euclidienne entre ces deux points est la longueur de l'hypothénuse du triangle rectangle. Il est également possible de prendre le carré de la distance euclidienne en enlevant la racine carrée de la formule ci-dessus. C'est une approche qui facilite le calcul et qui peut être utilisée, par exemple, dans la méthode de Ward (voir ci-après).

D'autres mesures de distance peuvent s'avérer appropriées dans le cadre de données métriques :

- **le coefficient de corrélation de Pearson** : c'est une mesure d'association qui permet d'établir si deux variales mesurées sur le même ensemble d'observations varient de façon analogue ou non ;
- **la distance de Tchebycheff** : il s'agit de la différence maximale absolue entre les valeurs relatives aux éléments de la classification ;
- **la distance de Minkowski** : c'est la racine *n*ième de la somme des différences absolues entre les valeurs relatives aux éléments à la puissance *n*.

Dans le cadre d'une classification avec des données binaires, on privilégiera :

- **l'indice de Sokal et Michener** : il représente le rapport entre les appariements (rapprochements deux à deux) et le nombre total de valeurs ;
- **l'indice de Rogers et Tanimoto** : cet indice attribue un poids deux fois plus important aux non-appariements (non-coïncidences) ;
- **l'indice de Sokal et Sneath** : un poids plus important est accordé aux appariements qui comptent le double.

De nombreuses autres mesures de distance existent ; il est fortement conseillé de tester empiriquement ces techniques afin de définir la mesure qui représentera de la manière la plus efficace la structure des données collectées.

Un certain nombre de limites doivent cependant être prises en compte. Lors d'analyses typologiques reposant sur des mesures différentes – par exemple des échelles de Likert, des pourcentages, des montants en euros, etc. –, il est nécessaire de standardiser les mesures et d'élimer les observations aberrantes. En ce qui concerne la standardisation, l'approche la plus courante est la méthode de l'écart type, mais d'autres approches peuvent être testées. L'utilisation de mesures de distance différentes peut conduire à des résultats de classification différents. L'analyse typologique est en ce sens une méthode empirique, où, comme nous l'avons signalé, de nombreuses combinaisons doivent être testées avant de déterminer la configuration optimale.

2.2 Constitution des groupes

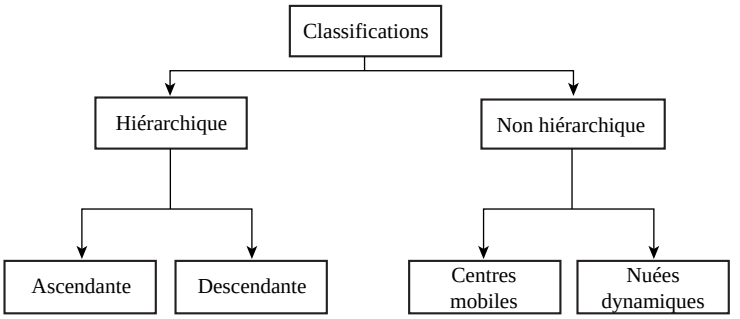
Il existe deux types de méthodes de constitution des groupes (classification) : les méthodes dites **hiérarchiques** et les méthodes **non hiérarchiques** (voir figure 4.4).

Les méthodes de **classification hiérarchique** consistent à établir une structure arborescente **ascendante** (à partir de chaque individu de groupe différent en constituant des groupes de plus en plus gros) ou **descendante** (à partir de tous les individus regroupés). Les méthodes de **classification non hiérarchique** visent à constituer *k* groupes (*k* étant

spécifié en début d'analyse) à partir des *n* individus de départ. Nous présentons dans cette section les méthodes les plus couramment mises en œuvre.

Figure 4.4

Choisir une méthode de classification.



Classification hiérarchique ascendante

La classification hiérarchique ascendante est un processus relativement simple et répétitif. Les individus/observations sont regroupés en segments aux caractéristiques communes. On peut définir le processus de classification comme suit :

- chaque observation représente un groupe, le nombre de groupes est par conséquent égal au nombre d'observations ;
- les deux groupes aux caractéristiques les plus proches sont agrégés au sein d'un même groupe en fonction de la mesure de distance choisie (euclidienne par exemple) et de la méthode d'agrégation retenue (voir ci-après) ;
- le processus d'agrégation ci-dessus est répété $n - 1$ fois (n étant le nombre d'observations), c'est-à-dire jusqu'à ce qu'il n'y ait plus qu'un seul groupe.

Prenons l'exemple d'une enquête comportant 100 observations : la classification démarre avec 100 groupes d'une observation, ensuite les deux groupes les plus proches sont agrégés, puis l'on recherche parmi les 99 groupes les deux groupes les plus proches, jusqu'à ce que les deux derniers groupes soient agrégés au sein d'un même et dernier groupe.

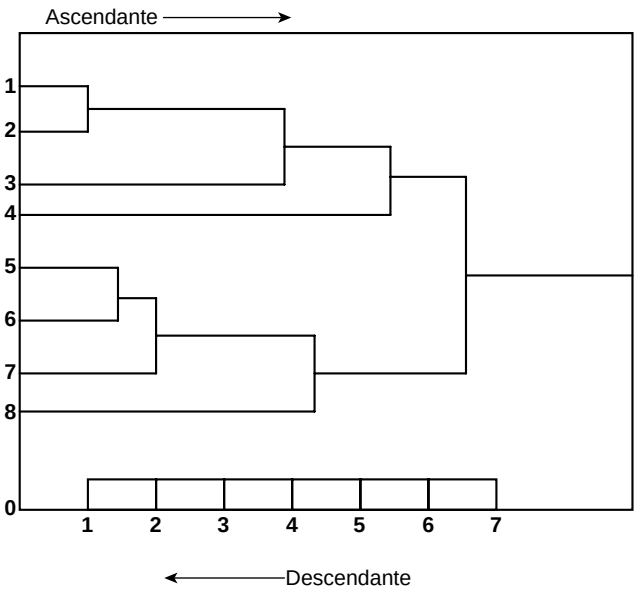
En ce qui concerne la constitution des groupes, là encore plusieurs approches peuvent être envisagées (Malhotra *et al.,* 2007). On retiendra cinq **méthodes** (ou algorithmes) **d'agrégation** principales :

- le **saut minimum** : cette méthode consiste à prendre la plus petite distance mesurée entre un élément de chaque groupe, puis la plus petite distance suivante, etc. ;
- la **distance du diamètre** : la distance entre deux groupes est calculée partir de la distance entre leurs deux points les plus éloignés ;
- la **distance moyenne** : cette méthode est relativement proche ; la distance entre deux groupes est définie par la moyenne des distances entre toutes les paires d'individus en prenant en compte, pour chaque paire, un membre de chaque groupe. Cette méthode est couramment employée du fait qu'elle utilise l'information de toutes les paires de distances ;
- la **méthode des barycentres** : il s'agit, comme la méthode de Ward, d'une méthode fondée sur la variance. Ces méthodes ont pour objet en effet de générer des groupes afin de minimiser la variance à l'intérieur de ceux-ci. On mesure la distance entre deux groupes en établissant la distance entre leurs barycentres (point construit à partir des moyennes de toutes les variables) ;

- la **méthode de Ward** : on calcule les moyennes pour toutes les variables de chaque groupe, puis, pour chaque individu, le carré de la distance euclidienne au centre de la classe.

Figure 4.5

Interprétation des deux grandes méthodes de classification.



La figure 4.5 représente ce que l'on nomme un **dendogramme**, autrement dit la représentation graphique des résultats de la classification des individus en groupes. On lit le dendogramme de gauche à droite pour une classification hiérarchique ascendante, et de droite à gauche pour une classification hiérarchique descendante. Les traits horizontaux de la partie gauche du dendogramme représentent les 8 segments finaux réalisés lors de la classification hiérarchique. La longueur de ces traits horizontaux est également caractéristique de la distance qui sépare les groupes. Les lignes verticales représentent l'agrégation, le rapprochement de deux groupes. Sur le graphique présenté, les segments 1 et 2 peuvent être agrégés (ils sont proches à une distance de 1 à peu près) [1], ce qui est également le cas des segments 5 et 6 (à une distance de 1,5 environ). L'agrégation suivante se fait à une distance de 2 et concerne le nouveau segment (composé des segments initiaux 5 et 6) et du segment 7. Nous développons l'interprétation d'un dendogramme plus en détail dans la partie suivante.

Classification hiérarchique descendante

La classification hiérarchique descendante consiste à considérer l'ensemble des observations rassemblées au sein d'un même segment, puis à les diviser en deux segments, puis en trois, quatre, etc., jusqu'à obtenir un nombre maximum de segments (des groupes ne contenant qu'un seul individu).

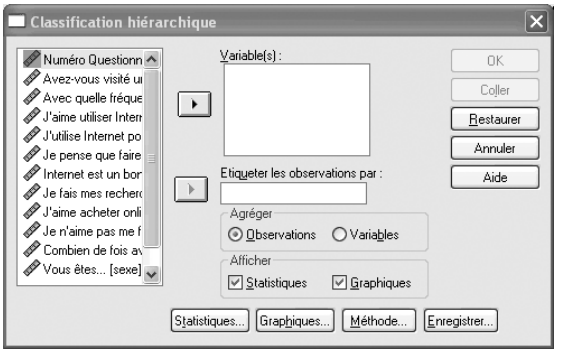
1. Les distances ici sont fictives, elles ont pour but d'illustrer l'écart relatif entre les groupes.

SPSS

La procédure à suivre dans SPSS est la suivante : Menu **Analyse > Classement > Classification hiérarchique…** La boîte de dialogue de la figure 4.6 s'affiche.

Figure 4.6

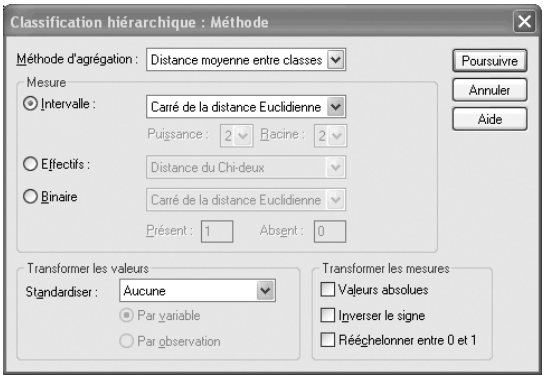
Boîte de dialogue du menu Classification Hiérarchique.



Si vous classez des observations, vous devez sélectionner au moins une variable numérique. Si vous classez des variables, sélectionnez au moins trois variables numériques. Il est également possible de sélectionner une variable d'information pour étiqueter les observations (par exemple classer les observations par pays). Le menu graphique vous permet de générer un **arbre hiérarchique** ou **dendogramme** (voir ci-après) souvent utile pour l'interprétation. En ce qui concerne la méthode (voir figure 4.7), plusieurs possibilités existent : le saut minimum, la distance du diamètre, la distance moyenne, la méthode des barycentres et la méthode de Ward (la plus couramment utilisée).

Figure 4.7

Boîte de dialogue du sous-menu Méthode.



Une fois la méthode retenue, la mesure doit être sélectionnée : l'analyste peut choisir entre la distance euclidienne ou le carré de la distance euclidienne par exemple, mais encore la corrélation de Pearson, la distance de Tchebycheff, la distance de Minkowski, l'indice de Sokal et Michener, l'indice de Rogers et Tanimoto, l'indice de Sokal et Sneath que nous avons abordés dans la section précédente. Il est préférable de tester plusieurs méthodes et plusieurs mesures avant de retenir une solution définitive. En effet, ces approches peuvent produire des résultats différents, plus ou moins lisibles ou utiles pour la décision. L'arbre de décision ou dendogramme peut faciliter la lecture des résultats.

2.3 Classification non hiérarchique

Les méthodes de classification non hiérarchiques visent à constituer *k* groupes (*k* étant spécifié en début d'analyse) à partir des *n* individus de départ. Ces méthodes sont très largement utilisées car elles permettent de traiter des volumes importants tout en optimisant les critères de classification. La particularité de ces méthodes, à la différence des méthodes de classification hiérarchique, est que le choix du nombre de groupes se fait en début de processus. Il s'agit d'un paramètre que l'analyste doit fixer avant de lancer l'algorithme. C'est un élément qui peut poser problème, étant donné que l'on ne connaît jamais *ex ante* le nombre idéal de groupes existant au sein d'une population étudiée. Dans la pratique, il est préférable de ne pas avoir à traiter un nombre trop élevé de groupes, ce qui rendrait difficile l'interprétation. On recommande donc de faire plusieurs essais avec des nombres de groupes différents afin d'identifier la meilleure solution au regard de critères statistiques de validité (la variance intergroupe divisée par la variance totale par exemple). On utilise en général des solutions comprenant entre 5 et 10 groupes.

On distingue deux méthodes principales de classification non hiérarchique, qui sont en réalité deux niveaux d'une même approche : la méthode des **centres mobiles** et la méthode des **nuées dynamiques**.

- **la méthode des centres mobiles** : méthode décomposant un ensemble d'individus en un nombre *n* de classes choisies *a priori* par un processus itératif convergeant de sélection des représentants de chaque classe (un centre par classe), qui peut être initialisé au hasard ou par l'utilisateur de la méthode. Les individus sont donc regroupés autour de ces centres de classe, les groupes étant constitués des individus les plus proches du centre du groupe. Une fois les individus affectés, on remplace les centres par les barycentres (c'est-à-dire le point d'équilibre de tous les points pris en compte) afin de recalculer les classes ;
- **la méthode des nuées dynamiques** : il s'agit d'une généralisation de la méthode des centres mobiles, dans laquelle chaque classe est représentée par un noyau de plusieurs éléments et non plus par un seul. Par ailleurs, le barycentre de chaque groupe est recalculé à chaque nouvel individu et non lors de l'affectation de tous les individus. La convergence est ainsi plus rapide et parfois même possible en une seule itération, ce qui peut être utile sur de gros volumes de données (Tufféry, 2006).

La méthode des nuées dynamiques est une méthode fréquemment employée. Elle est particulièrement performante en marketing, où le nombre d'observations (questionnaires collectés par exemple) est fréquemment supérieur à 100. Les classifications hiérarchiques sont en effet lourdes à manipuler au-delà de ce seuil, le nombre d'itérations étant trop important en termes de capacité de calcul. L'approche des nuées dynamiques est intéressante également car elle permet d'adopter un raisonnement utile pour l'interprétation. Elle suppose, en effet, qu'il existe pour chaque groupe un individu plus représentatif que les autres de la classe à laquelle il est affecté, celui qui est le plus proche du barycentre. Notons que les variables doivent être quantitatives et/ou que la mesure de distance employée est la distance euclidienne simple. Si vous souhaitez utiliser une autre mesure de distance, il est préférable d'utiliser la méthode de classification hiérarchique.

Focus 4.1 Application aux données textuelles

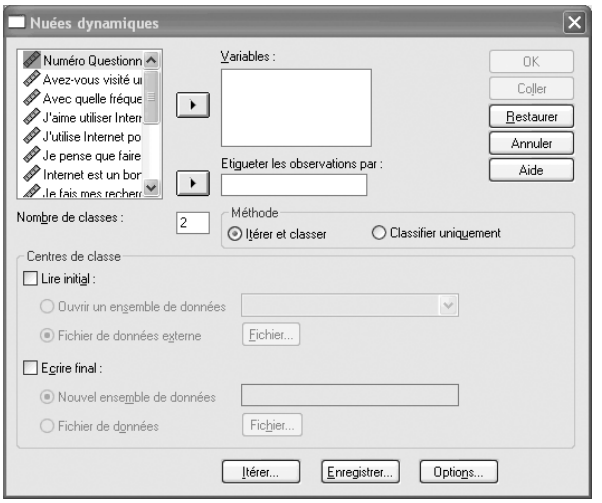
Les méthodes de classification hiérarchiques, mais également les nuées dynamiques, ont été depuis les travaux de Benzécri, puis de Reinert, appliquées au domaine de la lexicométrie (ou statistique textuelle), et plus récemment au Text Mining (extension aux données textuelles du Data Mining classique). Le principe sous-jacent au déploiement de ces méthodes est qu'il est possible de mettre à jour dans un discours ou un texte, une structure, des séquences qui vont permettre une analyse fine de ce type de données complexes à traiter. Plus précisément, la statistique textuelle a pour objet de découper un texte en unités textuelles (des mots par exemple) puis de regrouper les unités qui sont proches de façon à obtenir des classes homogènes de discours, suffisamment distinctes les unes des autres, que Reinert nomme des « mondes lexicaux » et qui correspondent aux différentes facettes d'un corpus textuel. Il est ainsi possible, par exemple, d'extraire de grandes classes de mots utilisés par les consommateurs pour parler d'une marque ; un premier groupe de mots fait référence à l'image, un deuxième au rapport qualité/prix, un troisième aux concurrents, etc. De nombreux logiciels permettent de traiter ces données textuelles : Alceste (le logiciel développé par Max Reinert du CNRS), Tropes, Sphinx Lexica, mais également dans une approche de Text Mining : Lexiquest et Clementine de SPSS et Text Miner de SAS.

SPSS

La procédure à suivre dans SPSS est la suivante : **Analyse > Classification > Nuées dynamiques…** La boîte de dialogue de la figure 4.8 apparaît.

Figure 4.8

Boîte de dialogue du menu Nuées dynamiques.



La première étape consiste à sélectionner les variables qui paraissent les plus pertinentes. On spécifie ensuite le nombre de classes que l'on souhaite obtenir (entre 5 et 10). Une indication du nombre de classes souhaitable peut être fournie par une première analyse de type ACP (analyse en composantes principales), par exemple pour simplifier des données collectées (voir le chapitre 3 sur la simplification des données). Le menu **Options** permet de spécifier un certain nombre d'éléments qui seront utiles à l'interprétation : préciser les centres de classe initiaux ou créer un tableau ANOVA afin de déterminer les variables les plus discriminantes dans la constitution des groupes et éliminer ainsi les centres de classe initiaux (pour l'interprétation de l'ANOVA, voir le chapitre 5). Il est possible également d'exclure les valeurs manquantes. On clique ensuite sur **Itérer** pour lancer la procédure.

3 Mise en œuvre

L'analyse typologique est une méthode qui suppose de tester empiriquement un grand nombre de combinaisons différentes. La nature des données à segmenter mais également les choix opérés au niveau de la mesure de distance et des méthodes de constitution des groupes rendent extrêmement complexe le choix d'une combinaison optimale. À titre d'exemple, le nombre de regroupements possibles de 1 000 personnes en 6 classes est de l'ordre de 10^{15}! (Evrard *et al.*, 1997) Dès lors, un certain nombre de problèmes pratiques se posent à l'analyste. Les deux premiers sont liés à la mise en œuvre de la démarche : le choix du nombre des groupes et l'interprétation du profil des groupes constitués. Le troisième problème est lié à la validité de la classification ; d'importantes précautions doivent être prises au cours de cette étape, comme nous le montre l'exemple suivant.

EXEMPLE

Pour illustrer la mise en œuvre d'une démarche de classification, prenons un exemple concret. Une entreprise du secteur informatique, fabriquant et commercialisant des ordinateurs et des baladeurs numériques, souhaite se diversifier en lançant un téléphone portable nouvelle génération. Elle réalise une étude de marché afin de lancer une gamme de produits déclinables en fonction de segments de consommateurs ayant des besoins suffisamment différenciés les uns des autres pour éviter toute cannibalisation. L'enquête a été administrée à 160 consommateurs regroupés en 7 segments différents et qui ont été interrogés sur leurs préférences, notées sur une échelle de 1 à 7, sur un total de 15 attributs :

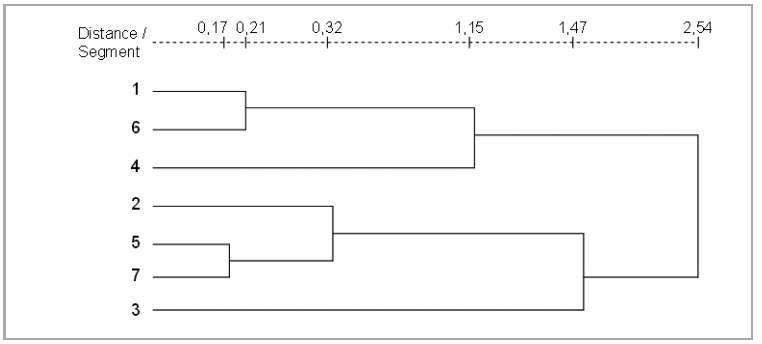
Intérêt nouveauté	Émission données	Fonction Internet
Utilisation SMS	Bluetooth	Appareil photo
Utilisation voix	Wi-Fi	Design
Utilisation agenda	Taille écran	Prix abonnement
Réception données	Fonction E-mail	Prix achat (hors abon.)

3.1 CHOISIR LE NOMBRE DE GROUPES

Étant donné le nombre d'observations, le chargé d'étude décide de mettre en place une classification hiérarchique afin de constituer les groupes. Les résultats sont représentés dans l'arbre de décision de la figure 4.9.

Figure 4.9

Représentation graphique des résultats de la première classification.



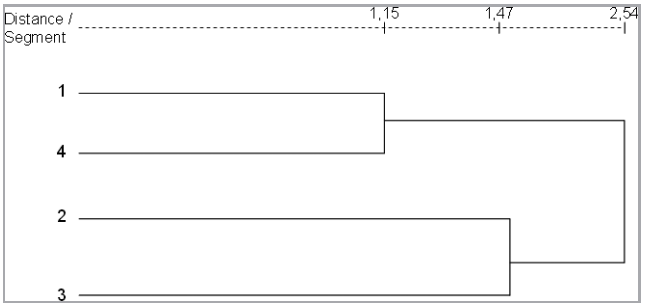
Le dendogramme nous fournit à la fois une visualisation graphique des résultats et le niveau en termes de distance des regroupements effectués. On peut voir, par exemple, que les segments 5 et 7 sont les plus proches, à une distance de 0,17 seulement. On constate ensuite que les deux segments suivants, les segments 1 et 6, ne sont séparés que par une distance de 0,21. Le « saut » suivant est effectué à une distance de 0,32 et regroupe l'agrégation des segments 5 et 7 avec le segment 2. On entend par « saut » les écarts de distance entre les regroupements effectués. On peut les identifier avec SPSS dans le fichier des résultats, où on les retrouve dans la **chaîne des agrégations**, tableau qui reprend les distances auxquelles ont été effectués les regroupements (voir exercice 1). On constate dans cet exemple que le saut suivant se situe à une distance de 1,15, qui correspond pratiquement au triple en termes de distance du précédent regroupement. Il existe donc un écart important entre les trois premiers regroupements et les suivants. Une solution à 4 classes semble donc pertinente (les 7 classes sont obtenues par 6 regroupements successifs ; si l'on fusionne les trois premiers regroupements énoncés, on n'obtient plus que 3 regroupements séparant 4 classes différentes).

3.2 Interpréter les groupes

Une lecture de la classification à 4 groupes donnerait les résultats présentés à la figure 4.10.

Figure 4.10

Représentation graphique la classification en 4 groupes.



La première étape dans l'interprétation de la classification obtenue consiste à revenir sur les **centres de groupes**. Plus précisément, on cherche à établir les coordonnées de ces points, que l'on pourrait assimiler à des centres de gravité des classes constituées, en reprenant les moyennes des scores des variables pour tous les individus appartenant à la classe. Il est important d'obtenir une classification pertinente du nombre de classes à exploiter mais également une lecture aisée de ces groupes d'individus ou de variables (voir tableau 4.1).

On décrit les segments obtenus en observant les scores moyens par variable et par groupe et en les comparant au score moyen de l'ensemble des répondants (colonne Total). On constate que le segment 1 est caractérisé par un intérêt fort porté à la nouveauté proposée par l'entreprise, par un bloc de variables (de Ut_Tel à Émission) correspondant aux fonctions classiques du téléphone et aux fonctions avancées (E-mail, Internet, Appareil photo). Le segment 2 est plutôt caractérisé par l'emploi des SMS, les éléments liés à la connectivité à distance, une taille d'écran importante. Le segment 3 peut être décrit par une forte émission/réception de données, un intérêt pour les éléments de connectivité à distance ainsi que pour la taille de l'écran, le design du produit ; il est relativement peu sensible au prix. Le segment 4 regroupe des individus attirés par la nouveauté, utilisant

Tableau 4.1 : Centres de groupes

Variable	Total	S 1	S 2	S 3	S 4
Intérêt	3,47	3,71	2,43	2,19	5,11
Ut_SMS	4,21	3,68	5,63	3,19	3,49
Ut_Tel	5,56	5,84	5,43	4,31	5,84
Ut_Agenda	4,01	5,89	2,33	3,06	3,86
Réception	4,45	5,02	3,88	6,12	3,65
Émission	4,50	5,20	3,90	6,25	3,51
Bluetooth	3,99	3,86	5,04	5,31	2,16
Wi-Fi	3,71	3,39	3,73	6,12	3,14
Écran	4,79	4,29	5,55	5,00	4,43
E-mail	4,72	5,96	3,31	2,88	5,59
Internet	4,47	5,66	3,04	1,44	5,97
Ap_Photo	4,01	5,20	5,45	1,94	5,27
Design	4,63	3,95	4,16	5,50	5,95
Px_Abon	28,8	24,6	25,3	45,3	32,6
Px_Achat	332	290	273	488	411

fortement leur téléphone, appréciant toutes les nouveautés technologiques proposées dans le nouveau produit et insensibles au prix. On peut considérer (on le voit également d'un point de vue graphique) qu'il existe deux segments principaux composés chacun de deux sous-segments. Les segments 1 et 4 regroupent en effet des individus attirés par la nouveauté. Le segment 4 étant moins sensible au prix, on pourrait les comparer à des *early adopters* ou adoptants précoces, qui sont les premiers à acheter les nouveautés sur le marché. Le segment 4 représente des individus attirés par la nouveauté mais relativement peu informés et relativement désargentés. Ils pourraient constituer une cible intéressante pour une seconde vie du produit, avec abonnement, une fois que l'innovation aura été diffusée auprès d'une première couche de population, plus rentable. Le second groupe, constitué des segments 2 et 3, représente une population qui diffère légèrement de la première. Le segment 3, caractérisé par les variables depuis Émission jusqu'à Écran, est sensible au design et très peu au prix. Il s'agit d'un segment probablement CSP + ou professionnel, à qui l'on peut destiner une version haut de gamme, tant d'un point de vue technique qu'en ce qui concerne les services associés (ils téléchargent et émettent un volume important de données). Le segment 2 est un segment plus *mass market a priori*, qui pourrait correspondre à une population plus jeune (SMS), connectée (Bluetooth, Wi-Fi) et qui souhaite uti-

liser les fonctionnalités multimédias de l'appareil (Écran, Appareil photo) afin de communiquer.

Pour s'assurer de la validité de la classification obtenue, il est recommandé de vérifier en premier lieu la cohérence au sein des différents groupes (effectuer une analyse de variance par exemple). L'analyste peut également réaliser des tests statistiques sur chaque variable (fréquences, etc.) afin de comparer les résultats au sein d'un groupe avec l'ensemble des observations. Ces démarches ne sont utiles que si un certain nombre de combinaisons (méthode, distance, ajout/omission de variables, etc.) ont déjà été testées.

Résumé

L'analyse typologique est une méthode fréquemment mobilisée en analyse de données. Elle permet non seulement de classer des individus ou des variables, mais également de réduire les données en les regroupant au sein de classes homogènes. Il n'existe pas une mais des méthodes de segmentation. Que l'on opte pour une procédure de classification hiérarchique ou non hiérarchique, l'analyse typologique confère une grande liberté à l'analyste, mais rend également plus complexe le choix de la bonne approche. Elle suppose de tester empiriquement un grand nombre de combinaisons avant de trouver la démarche qui aboutisse à des résultats exploitables (nombre de groupes et interprétation) et valides.

Pour aller plus loin

Evrard Y., Pras B., Roux E., *Market. Études et recherche en marketing*, Nathan, Paris, 2003.

Hair J. F., Anderson R. E., Tatham R. L., Black W. C., *Multivariate Data Analysis*, Prentice Hall International, New Jersey, 2007.

Malhotra N., Decaudin J. M., Bouguerra A., *Études marketing avec SPSS*, 5e éd., Pearson Education, Paris, 2007.

Tufféry S., *Data mining et statistiques décisionnelles*, éditions Technip, Paris, 2007.

Exercices

Exercice 1 Habitudes alimentaires

Énoncé

Une enseigne de grande distribution cherche à réaliser une enquête sur les habitudes alimentaires en Europe afin d'adapter sa politique d'achat et de référencement. Les données concernant 25 pays ont été recueillies. Elles portent sur les indices globaux de consommation de 9 catégories de produits alimentaires : viande rouge, viande blanche, œuf, lait, poisson, céréales, féculents, oléagineux, fruits et légumes. Les données issues de l'enquête sont disponibles dans le fichier « alimentaire.sav ».

1. Quelle pourrait être l'utilité de l'analyse typologique dans ce cas précis ?
2. Quelle méthode de classification recommandez-vous ?
3. Réalisez et décrivez l'arbre de classification.
4. Dans le cas de la classification hiérarchique, on peut également interpréter le nombre de groupes par le biais de la chaîne d'agrégation, qui reprend dans un tableau les distances auxquelles les groupes sont agrégés. L'interprétation de cette chaîne consiste à repérer des « sauts » de distance dans la constitution des groupes. Commentez le tableau de la chaîne d'agrégation obtenu.
5. Combien de groupes faut-il garder ?
6. Êtes-vous satisfait des résultats de l'analyse ?

Solution

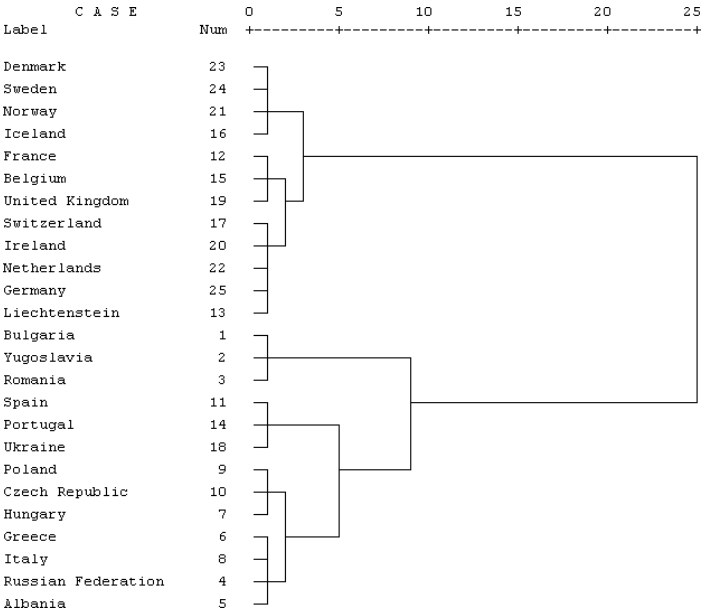
1. L'analyse typologique permet de « réduire le nombre d'observations en les regroupant en des classes homogènes et différenciées ». Dans ce cas précis, l'objectif de l'analyse typologique pourrait être de faire apparaître des catégories de pays en fonction des habitudes alimentaires. L'utilité pour l'enseigne est multiple : en faisant émerger ces grands types de consommation, elle sera à même d'optimiser sa stratégie de négociation avec les fournisseurs, sa politique d'achat, structurer son département achat par régions, etc.

2. L'enquête porte sur seulement 25 observations, une classification hiérarchique semble pertinente. Dans ce cas précis, rien ne nous oriente vers une classification hiérarchique ascendante ou descendante. Il est préférable de s'orienter vers les pratiques les plus diffusées : nous pourrions essayer dans un premier temps de réaliser une classification hiérarchique ascendante. Un premier essai en utilisant la méthode de Ward et le carré de la distance euclidienne (préférable lorsqu'on utilise la méthode de Ward comme nous l'avons vu) pourrait s'avérer fructueux.

3. La procédure est la suivante : **Analyse > Classement > Classification hiérarchique…** Faites glisser les variables de « viande rouge » à « fruits et légumes » dans la case **Variable(s)**, et sélectionnez « Nom du pays » afin d'étiqueter les observations. Dans le menu **Graphique** sélectionnez **Arbre hiérarchique**. En ce qui concerne la méthode, on peut, dans un premier temps, essayer d'utiliser la méthode de Ward combinée à une mesure par le carré de la distance euclidienne. On obtient le dendogramme de la figure 4.11.

Figure 4.11

Représentation graphique des résultats de la première classification.



On peut observer sur l'arbre de décision que deux grands blocs de pays se détachent assez nettement. La première classe est constituée des pays allant du Danemark au Liechtenstein, la seconde de la Bulgarie à l'Albanie. Ces deux blocs sont repris dans le tableau 4.2.

Tableau 4.2 : Les deux premiers groupes de pays

Groupe 1	Groupe 2
Danemark	Bulgarie
Suède	Yougoslavie
Norvège	Roumanie
Islande	Espagne
France	Portugal
Belgique	Ukraine
Royaume-Uni	Pologne
Suisse	République tchèque
Irlande	Hongrie
Pays-Bas	Grèce
Allemagne	Italie
Liechtenstein	Russie
	Albanie

La classification semble assez cohérente. Le premier groupe correspond *a priori* à des pays plus développés, ou faisant partie du premier cercle de l'Union européenne d'un point de vue historique. Le second groupe, à l'exception de l'Italie, fait partie de pays ayant adhéré plus tardivement à l'UE ou hors UE. On peut supposer que, même si les écarts de développement ont été rattrapés pour certains d'entre eux (Espagne, Grèce, etc.), les difficultés

passées se notent dans les comportements alimentaires. Les données dont nous disposons ne nous permettent pas encore de véritable interprétation.

4. En ce qui concerne la chaîne d'agrégation, on obtient les résultats de la figure 4.12.

Figure 4.12

Chaîne d'agrégation de la typologie pays.

	Regroupement de classes			Etape d'apparition de la classe		
Etape	Classe 1	Classe 2	Coefficients	Classe 1	Classe 2	Etape suivante
1	23	24	11,500	0	0	6
2	1	2	23,385	0	0	10
3	17	20	36,375	0	0	14
4	12	15	54,420	0	0	7
5	22	25	75,710	0	0	13
6	21	23	97,083	0	1	17
7	12	19	122,892	4	0	19
8	6	8	154,707	0	0	12
9	9	10	188,782	0	0	15
10	1	3	223,897	2	0	23
11	11	14	262,517	0	0	18
12	4	6	304,035	0	8	16
13	13	22	346,405	0	5	14
14	13	17	399,447	13	3	19
15	7	9	464,119	0	9	20
16	4	5	541,265	12	0	20
17	16	21	640,170	0	6	21
18	11	18	750,156	11	0	22
19	12	13	866,915	7	14	21
20	4	7	1034,679	16	15	22
21	12	16	1269,050	19	17	24
22	4	11	1766,153	20	18	23
23	1	4	2632,676	10	22	24
24	1	12	5243,414	23	21	0

Nous recherchons des « sauts » de distance dans la chaîne d'agrégation. Le premier saut apparaît nettement et confirme la description en deux classes principales : la distance double entre les étapes 23 et 24 (de 2 632,676 et 5 243,414). Le deuxième saut (22-23) est caractérisé par un écart de 900 environ, le troisième saut (21-22) par un écart de 500 environ, et enfin le quatrième saut (20-21) par un écart de 200 seulement. Les troisième et quatrième sauts étant caractérisés par des écarts trop faibles si on les compare aux deux précédents, une solution à trois classes semble se profiler. Cette solution nous amènera à distinguer deux sous-groupes au sein du deuxième groupe de pays : un groupe constitué de la Bulgarie, de la Yougoslavie et de la Roumanie, d'un côté (ce qui a du sens d'un point de vue purement géographique) et le reste des pays, de l'autre.

5. Au vu des résultats précédents, et des objectifs que pourrait potentiellement mettre en œuvre l'enseigne de distribution, il semble qu'une solution à deux groupes soit préférable. En effet, le troisième groupe que nous avons fait apparaître n'étant constitué que de trois pays, la portée managériale de cette distinction est faible (mettre en place une cellule ou adapter la stratégie pour ces trois pays). Il faudrait croiser l'analyse avec d'autres variables, de type risque pays par exemple, qui sont fournies par les grands organismes internationaux (FMI, Banque mondiale, OMC, Eurostat, etc.) pour savoir s'il s'agit de pays à exclure des décisions stratégiques dans cette région.

6. Ces commentaires sont effectués sur les résultats d'une seule analyse. Ils ne donnent pas entière satisfaction et il est souhaitable de tester d'autres approches avant de donner un résultat définitif. À vous de tester d'autres procédures pour mieux déterminer les groupes.

Exercice 2 Achats On-line

Énoncé

Une enquête portant sur un nombre élevé de répondants (1 400 questionnaires exploitables) vient d'être réalisée. L'objet de cette enquête, commanditée par une chaîne de magasins spécialisée dans l'électroménager est de mieux comprendre le comportement multicanal du consommateur, c'est-à-dire si son comportement on-line diffère de son comportement off-line (en magasin traditionnel). Une première approche en termes d'analyse des résultats est de faire émerger des types de répondants. Une extraction des résultats de cette enquête est disponible dans le fichier « on-line.sav » disponible sur le site : http://www.pearsoneducation.fr.

1. Quelle démarche peut-on mettre en œuvre ? Argumentez.
2. Décrivez puis interprétez les segments obtenus.

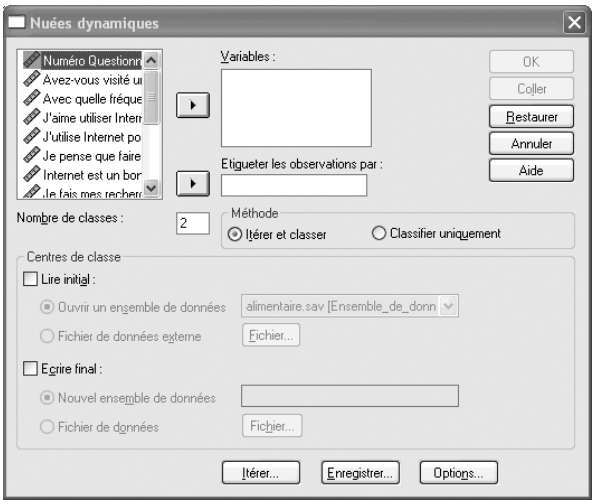
Solution

1. Le nombre élevé d'observations nous oriente assez naturellement vers une méthode de classification non hiérarchique. Ces méthodes, rappelons-le, visent à constituer *k* groupes (nombre spécifié dès le départ) à partir des *n* individus (1 400 dans cet exemple). Le choix d'une méthode non hiérarchique s'explique par le recours à un nombre moins élevé d'itérations que dans le cas d'une classification hiérarchique, ce qui « allège » l'algorithme en termes de capacité de calcul (si vous lancez SPSS avec une classification hiérarchique sur ces données vous risquez d'attendre très longtemps vos résultats !). Enfin, les méthodes non hiérarchiques que nous avons abordées (centre mobiles et nuées dynamiques) sont intéressantes en termes d'interprétation car elles supposent qu'il existe un centre de classe, c'est-à-dire un individu plus représentatif de son groupe d'appartenance. Il semble donc judicieux de mettre en œuvre une analyse par la méthode des nuées dynamiques (*K-means*).

2. Pour lancer la méthode des nuées dynamiques reprenez la démarche vue dans la partie cours : **Analyse > Classement > Nuées dynamiques…** La procédure affiche la boîte de dialogue de la figure 4.13.

Figure 4.13

Boîte de dialogue de la méthode des nuées dynamiques.



La première étape consiste à choisir les variables les plus adaptées à l'analyse. Vous pouvez vous aider des étiquettes des variables (dans l'éditeur de données cliquez sur l'onglet « affichage des variables »). Nous n'avons pas vraiment d'indication, en dehors des questions posées (pas d'analyse factorielle par exemple sur la structure des données). Nous pouvons inclure, dans un premier temps, l'ensemble des variables (à l'exception du numéro de questionnaire, sans objet). Faites glisser les variables dans la cellule « variable(s) ». Choisissez ensuite le nombre de classes que vous souhaitez obtenir : une AFC peut être utile ici pour vous orienter. Vous pouvez la réaliser en vous reportant au chapitre 3.

Nous allons procéder de manière plus empirique. Tout d'abord, nous choisissons un nombre légèrement plus élevé de classes que celui attendu *a priori* (ou suggéré par l'AFC/ACP). Les données que nous utilisons sont pour la plupart des échelles en 5 points, sauf la variable dichotomique sur la récence de la visite d'un site d'e-commerce qui pourrait être une variable relativement discriminante (de même que le sexe). Proposons dans un premier temps une classification en 4 classes et observons les résultats. Étant donné que nous allons classer un nombre élevé de variables, il faut augmenter le menu **Itérer** à 30 itérations maximum (nous pourrons augmenter/diminuer le nombre d'itérations si les résultats nous indiquent que ce nombre s'avère insuffisant/trop élevé). Il est possible, dans ce menu, de sauvegarder les classes en tant que nouvelles variables ; cette opération est utile en fin d'analyse pour vérifier la validité des résultats. Dans les **Options**, choisissez d'ajouter un tableau ANOVA (analyse de variance) qui sert à déterminer quelles sont les variables les plus discriminantes dans la constitution des groupes. Lancez la procédure. Nous obtenons les résultats présentés à la figure 4.14.

Figure 4.14

Historique des itérations.

Historique des itérations[a]

Itération	Changements dans les centres de classes			
	1	2	3	4
1	3,640	3,989	4,052	3,592
2	,339	,678	1,213	,180
3	,205	,301	,490	,114
4	,129	,147	,197	,081
5	,022	,065	,096	,084
6	,030	,048	,058	,046
7	,017	,015	,071	,036
8	,008	,007	,087	,047
9	,011	,011	,092	,060
10	,000	,016	,065	,052
11	,000	,018	,058	,054
12	,000	,014	,051	,037
13	,000	,012	,049	,040
14	,011	,022	,028	,032
15	,000	,024	,023	,032
16	,000	,025	,047	,053
17	,008	,037	,017	,039
18	,009	,014	,032	,031
19	,000	,009	,028	,024
20	,006	,006	,036	,025
21	,000	,000	,067	,048
22	,000	,010	,028	,027
23	,000	,000	,000	,000

a. La convergence obtenue est due à l'absence ou à la quasi-absence de modifications dans les centres de classes. La modification absolue maximale des coordonnées d'un centre est ,000. L'itération en cours est 23. La distance minimale entre les centres initiaux est 9,110.

Dans la plupart des cas on peut laisser le menu itérer par défaut (à 10 itérations maximum). Les classes convergent dans notre cas avant la 23e itération, c'est-à-dire que la valeur ,000 est atteinte dans chacune des quatre classes.

On observe ensuite le nombre d'observations dans chaque classe. Il faut veiller à ce que celles-ci ne soient pas trop déséquilibrées. Une bonne pratique consiste à ne garder que les classes qui représentent 10 % ou plus des observations. Dans notre cas, on obtient la répartition de la figure 4.15.

Figure 4.15

Nombre d'observations dans chaque classe.

Classe	1	375,000
	2	409,000
	3	261,000
	4	355,000
Valides		1400,000
Manquantes		,000

La répartition semble homogène. Si les résultats avaient été déséquilibrés à ce niveau, il aurait fallu éliminer une classe. Étudions maintenant les variables les plus discriminantes en décrivant les résultats du tableau ANOVA (voir figure 4.16).

Figure 4.16

Tableau ANOVA.

	Classe		Erreur			
	Moyenne des carrés	ddl	Moyenne des carrés	ddl	F	Signification
Avez-vous visité un site de commerce en ligne visités dans les 3 derniers mois?	,000	3	,000	1396	.	.
Avec quelle fréquence achetez-vous sur Internet?	625,445	3	,361	1396	1733,540	,000
J'aime utiliser Internet.	443,587	3	,646	1396	687,011	,000
J'utilise Internet pour rechercher les produits que souhaite acheter.	409,005	3	,724	1396	564,806	,000
Je pense que faire des achats sur Internet est sur.	603,145	3	,403	1396	1495,489	,000
Internet est un bon outil pour recherche un produit à acheter	355,857	3	,838	1396	424,771	,000
Je fais mes recherches online mais j'achète en magasin traditionnel.	1,769	3	1,024	1396	1,729	,159
J'aime acheter online.	1,900	3	1,388	1396	1,369	,251
Je n'aime pas me faire conseiller par un vendeur.	1,317	3	2,628	1396	,501	,682
Combien de fois avez-vous visité le site avant de faire votre achat?	4031,373	3	1,413	1396	2853,781	,000
Vous êtes...	4,013	3	,150	1396	26,735	,000

Les tests F ne doivent être utilisés que dans un but descriptif car les classes ont été choisies de manière à maximiser les différences entre les observations des diverses classes. Les niveaux de signification observés ne sont pas corrigés et ne peuvent par conséquent pas être interprétés comme des tests de l'hypothèse que les moyennes des classes sont égales.

Le test de significativité *F* est utilisé dans l'interprétation de l'analyse de variance (voir chapitre 4). Ici, le *F* ne doit être utilisé que dans un but descriptif car il s'agit de maximiser les différences entre les observations des différentes classes. On recherche seulement les valeurs significatives de *F* les plus élevées. Deux constats peuvent être faits à la lecture du tableau : les variables les plus discriminantes pour la constitution des classes sont : 1) les visites avant achat ($F = 2\,853{,}781$) ; 2) la fréquence d'achat sur Internet ($F = 1\,733{,}540$) ; 3) le sentiment de sécurité lors de l'achat en ligne ($F = 1\,495{,}489$). Le second constat provient des variables qui peuvent être éliminées de l'analyse : il s'agit des variables *vendeur* (« je n'aime pas me faire conseiller par un vendeur »), *on-line* (« j'aime acheter on-line ») et *multicanal* (« je fais mes recherches on-line mais j'achète en magasin traditionnel ») qui ne sont pas significatives (respectivement à 0,682/0,251/0,159). On peut relancer l'analyse en éliminant ces trois variables : nous obtenons alors une convergence en 20 itérations et les résultats présentés aux figures 4.17, 4.18 et 4.19.

On interprète les 4 classes en fonction des centres de classe finaux. On s'aperçoit assez rapidement qu'il s'agit d'hommes principalement et que les segments sont divisés en deux catégories principales : les pour et les contre (pour l'interprétation, on reprend la signification des valeurs en fonction des étiquettes de variables : 1 = absolument pas d'accord, etc.). On peut lancer une segmentation à deux classes pour faire apparaître plus clairement cette dichotomie. Les figures 4.20, 4.21, 4.22 et 4.23 présentent les résultats.

Figure 4.17

Résultats deuxième analyse par les nuées dynamiques (1).

Nombre d'observations dans chaque classe

Classe	1	408,000
	2	352,000
	3	264,000
	4	376,000
Valides		1400,000
Manquantes		,000

Figure 4.18

Résultats deuxième analyse par les nuées dynamiques (2).

	Classe		Erreur			
	Moyenne des carrés	ddl	Moyenne des carrés	ddl	F	Signification
Avez-vous visité un site de commerce en ilgne visités dans les 3 derniers mois?	,000	3	,000	1396	.	.
Avec quelle fréquence achetez-vous sur Internet?	625,598	3	,360	1396	1735,540	,000
J'aime utiliser Internet.	443,675	3	,645	1396	687,351	,000
J'utilise Internet pour rechercher les produits que souhaite acheter.	408,071	3	,726	1396	561,960	,000
Je pense que faire des achats sur Internet est sur.	603,305	3	,403	1396	1497,166	,000
Internet est un bon outil pour recherche un produit à acheter	356,145	3	,837	1396	425,431	,000
Combien de fois avez-vous visité le site avant de faire votre achat?	4032,267	3	1,411	1396	2858,303	,000
Vous êtes...	4,010	3	,150	1396	26,710	,000

Les tests F ne doivent être utilisés que dans un but descriptif car les classes ont été choisies de manière à maximiser les différences entre les observations des diverses classes. Les niveaux de signification observés ne sont pas corrigés et ne peuvent par conséquent pas être interprétés comme des tests de l'hypothèse que les moyennes des classes sont égales.

Figure 4.19

Résultats deuxième analyse par les nuées dynamiques (3).

	Classe			
	1	2	3	4
Avez-vous visité un site de commerce en ilgne visités dans les 3 derniers mois?	1	1	1	1
Avec quelle fréquence achetez-vous sur Internet?	2	4	4	1
J'aime utiliser Internet.	2	4	4	2
J'utilise Internet pour rechercher les produits que souhaite acheter.	3	4	4	2
Je pense que faire des achats sur Internet est sur.	2	4	4	1
Internet est un bon outil pour recherche un produit à acheter	2	3	4	2
Combien de fois avez-vous visité le site avant de faire votre achat?	6	8	11	3
Vous êtes...	1	1	1	1

Figure 4.20

Résultats finaux (1).

Itération	Changements dans les centres de classes	
	1	2
1	6,866	4,611
2	,505	,496
3	,209	,193
4	,097	,086
5	,036	,031
6	,000	,000

Figure 4.21

Résultats finaux (2).

Classe	1	644,000
	2	756,000
Valides		1400,000
Manquantes		,000

Figure 4.22

Résultats finaux (3).

	Classe		Erreur			
	Moyenne des carrés	ddl	Moyenne des carrés	ddl	F	Signification
Avez-vous visité un site de commerce en ilgne visités dans les 3 derniers mois?	,000	1	,000	1398	.	.
Avec quelle fréquence achetez-vous sur Internet?	1559,620	1	,587	1398	2657,731	,000
J'aime utiliser Internet.	1136,599	1	,784	1398	1450,413	,000
J'utilise Internet pour rechercher les produits que souhaite acheter.	987,451	1	,894	1398	1103,944	,000
Je pense que faire des achats sur Internet est sur.	1505,960	1	,620	1398	2429,714	,000
Internet est un bon outil pour recherche un produit à acheter	911,912	1	,948	1398	962,028	,000
Combien de fois avez-vous visité le site avant de faire votre achat?	9012,057	1	3,615	1398	2492,793	,000
Vous êtes...	11,399	1	,150	1398	75,813	,000

Figure 4.23

Résultats finaux (4).

	Classe	
	1	2
Avez-vous visité un site de commerce en ilgne visités dans les 3 derniers mois?	1	1
Avec quelle fréquence achetez-vous sur Internet?	4	2
J'aime utiliser Internet.	4	2
J'utilise Internet pour rechercher les produits que souhaite acheter.	4	2
Je pense que faire des achats sur Internet est sur.	4	2
Internet est un bon outil pour recherche un produit à acheter	4	2
Combien de fois avez-vous visité le site avant de faire votre achat?	9	4
Vous êtes...	1	1

On note tout d'abord que le calcul a été plus rapide (6 itérations seulement) et que les deux classes sont relativement homogènes (644 et 756 individus respectivement). Le premier groupe correspond à des habitués de l'achat en ligne, qui ont visité récemment un site d'e-commerce, qui achètent régulièrement sur Internet, et qui ont visité de nombreuses fois le site avant de faire leur achat, que ce soit sur le site ou en point de vente traditionnel. Le second segment correspond à des consommateurs qui sont plus réfractaires au commerce en ligne et qui ont visité peu de fois le site avant de faire leur achat. Dans les deux classes il s'agit principalement d'hommes. Nous n'avons pas d'information sur le canal utilisé pour réaliser l'achat : site ou magasin traditionnel. Une piste intéressante à suggérer à votre responsable : mettre en œuvre une analyse plus avancée pour expliquer l'achat on- ou off-line par l'ensemble des variables que nous venons d'étudier.

Exercice 3 Segmenter le marché automobile

Énoncé

L'exemple [1] que nous allons étudier reprend des informations sur les caractéristiques de différents modèles concurrents sur le marché US, ainsi que leur performance en termes de prix et de vente. L'objet de l'application est de réaliser une typologie des principales marques en présence sur ce marché. Ouvrez le fichier « ventes_voitures.sav » disponible sur le site : http://www.pearsoneducation.fr.

1. Peut-on, sur ces données, mettre en œuvre une classification hiérarchique ascendante ? Décrivez les étapes nécessaires à sa mise en œuvre.
2. Décrivez et interprétez les segments obtenus.

Solution

1. Nous allons procéder à une classification hiérarchique ascendante. Comme nous l'avons signalé, cette méthode est peu performante sur de gros volumes de données. Le tableau de données contenant 157 modèles concurrents, il est souhaitable de sélectionner les observations pour en retenir un nombre moins élevé. Nous pouvons centrer notre analyse sur les modèles les plus performants sur le marché par le biais de la procédure « sélectionner les observations » (que nous avons abordée au chapitre 2).

Dans le menu **Données** choisissez le sous-menu **Sélectionner les observations…** Nous nous intéressons aux modèles ayant vendu plus de 100 000 unités sur le marché américain. Sélectionnez les observations **selon une condition logique : Si** « (type = 0) & (ventes > 100) » comme indiqué sur la figure 4.24.

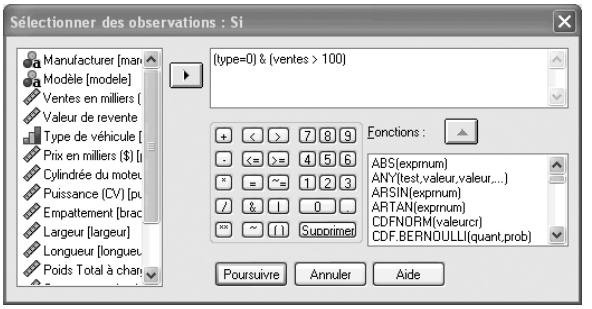
Pour lancer la classification hiérarchique ascendante, suivez les procédures que nous avons vues plus haut : **Analyse > Classement > Classification hiérarchique…**

Nous allons fonder notre analyse sur un certain nombre de variables de classification pertinentes dans le cas d'une segmentation de produits. Sélectionnez les variables allant de *Prix en millier ($)* jusqu'à *Consommation* en les faisant glisser dans la cellule « Variable(s) ». Nous allons utiliser une variable afin d'ordonner les résultats : faites glisser la variable *Modèle* dans la cellule « Etiqueter les observations par ». Dans le menu graphi-

1. Il s'agit ici d'une version légèrement modifiée d'un fichier disponible dans les fichiers exemples de SPSS (car_sales.sav). De nombreux fichiers de ce type peuvent être utilisés pour manipuler et expérimenter les différents tests disponibles sur SPSS.

Figure 4.24

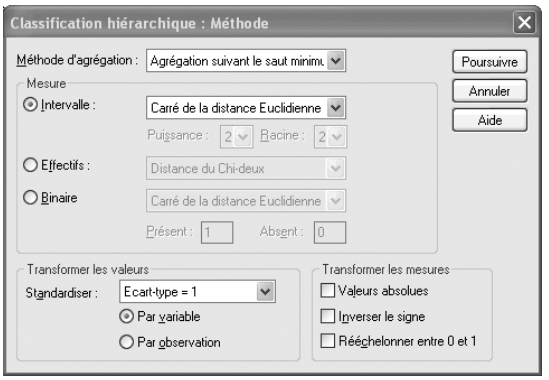
Boîte de dialogue Sélectionner les données selon une condition logique.



que, cochez la case « Arbre hiérarchique » et sélectionnez la case « Aucun » dans le sous-menu **Stalactites** afin de produire le dendogramme que nous analyserons dans la section suivante. Poursuivez et cliquez sur « Méthode » afin d'afficher la boîte de dialogue des mesures de distance de la classification. Nous allons procéder par une agrégation suivant le saut minimum, c'est-à-dire en déterminant la plus petite distance mesurée entre un élément de chaque groupe, puis la plus petite distance suivante, et ainsi de suite. Les données étant mesurées sur des échelles différentes (dollars, litres, etc.) nous allons les standardiser par l'emploi de l'écart type, comme indiqué sur la figure 4.25.

Figure 4.25

Boîte de dialogue Choix de la méthode d'agrégation et de la mesure de distance.

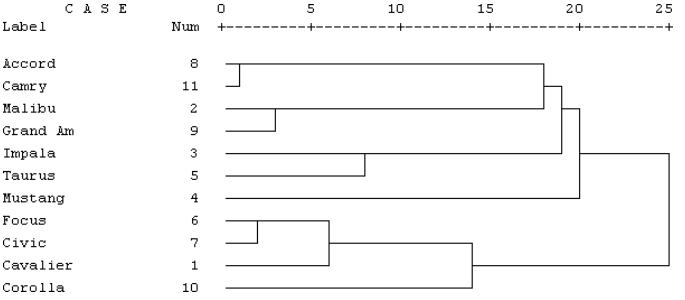


Cliquez sur **Poursuivre** et lancez la classification.

La classification hiérarchique ascendante nous a permis d'obtenir 10 segments de véhicules, comme le montre le dendogramme à la figure 4.26.

Figure 4.26

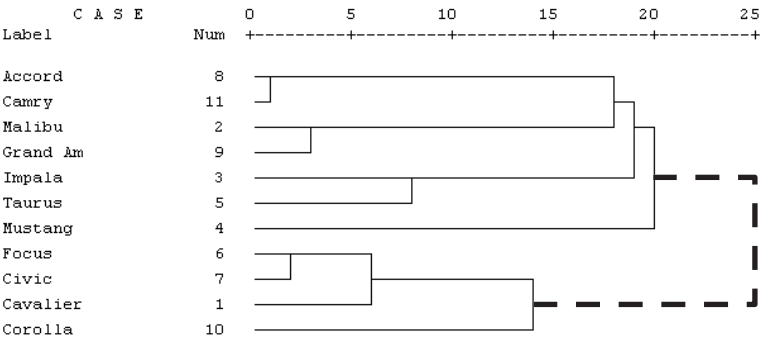
Dendogramme des résultats de la classification hiérarchique.



2. Lorsqu'on cherche à définir le nombre de groupes idéal sur la base d'un dendogramme, on souhaite identifier de manière empirique des « sauts » de distance entre les différents regroupements effectués. En lisant le dendogramme de la droite vers la gauche, on constate qu'il y a un saut important entre 25 et 20 qui sépare le marché automobile en deux segments principaux, comme le montrent les traits en pointillés sur la figure 4.27.

Figure 4.27

Lecture du premier segment sur le dendogramme.

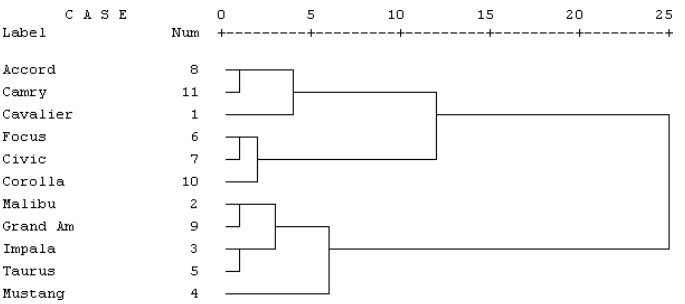


Rappelons la règle de lecture du dendogramme énoncée plus haut : les axes verticaux représentent les regroupements de segments, les axes horizontaux les distances entre les segments. Il s'agira donc de ne conserver que les segments les plus distants et d'agréger les segments les plus proches (ayant la plus petite distance). On peut constater, si l'on poursuit la lecture des résultats graphiques, qu'il existe un autre saut, entre 10 et 15, qui suggère 6 segments différents si l'on compte le nombre de lignes horizontales entre ces deux mesures. On peut encore lire les résultats différemment : on constate qu'il existe un écart visuel entre les 5 premiers axes verticaux (les 5 premiers regroupements suggérant donc 6 segments) et les axes verticaux suivants situés sur la partie gauche du graphique. En effet, le 5e axe se situe à une distance de 14 à peu près, tandis que le suivant se situe à une distance de 8.

Une telle solution à 6 segments n'apporte pas suffisamment de clarté à notre lecture du marché automobile. Il peut être intéressant alors d'essayer une autre méthode d'agrégation qui pourrait s'avérer plus pertinente. Rappelez la boîte de dialogue et choisissez la méthode d'agrégation suivant la distance du diamètre, calculée à partir de la distance entre les deux points les plus éloignés des groupes comparés deux à deux. On obtient l'arbre de classification présenté à la figure 4.28.

Figure 4.28

Lecture du second dendogramme.



On voit apparaître les résultats plus clairement. Deux segments différents peuvent être identifiés : les modèles du haut (de Accord à Corolla) représentent les véhicules les plus petits, les modèles du bas (de Malibu à Mustang) représentent les modèles les plus grands. On peut remarquer également que deux sous-segments se distinguent au sein des plus petits véhicules : la Focus, la Civic et la Corolla sont des véhicules moins chers que les trois modèles du haut.

En ce qui concerne la chaîne des agrégations de notre second cas, on obtient le tableau de la figure 4.29.

Figure 4.29

Chaîne des agrégations.

	Regroupement de classes			Etape d'apparition de la classe		
Etape	Classe 1	Classe 2	Coefficients	Classe 1	Classe 2	Etape suivante
1	8	11	1,260	0	0	7
2	6	7	1,579	0	0	5
3	2	9	1,625	0	0	6
4	3	5	2,619	0	0	6
5	6	10	4,012	2	0	9
6	2	3	7,333	3	4	8
7	1	8	9,183	0	1	9
8	2	4	12,440	6	0	10
9	1	6	25,486	7	5	10
10	1	2	54,607	9	8	0

Dans ce tableau, le coefficient d'agrégation réalise un saut important (plus du double) entre les étapes 9 et 10 : la solution à deux groupes est bien appropriée.

Chapitre 5

L'analyse de variance

L'analyse de variance désigne une famille de méthodes destinées à examiner et à interpréter les différences de moyennes observées entre plusieurs groupes pour une même variable (ANOVA pour *ANalysis Of VAriance*) ou pour plusieurs variables (MANOVA pour *Multivariate ANalysis Of VAriance*). Ces méthodes sont souvent utilisées pour analyser des données issues d'une expérimentation où des caractéristiques d'un objet sont manipulées afin d'optimiser l'objet au moindre coût.

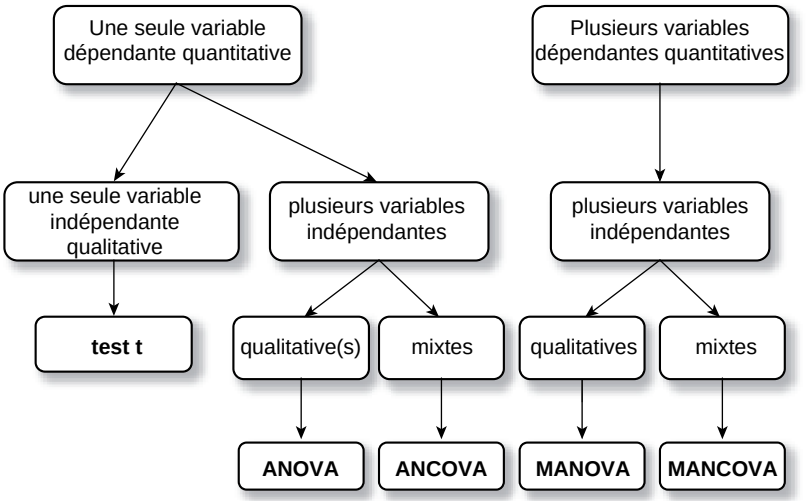
Nous verrons dans un premier temps les techniques d'analyse de variance et de covariance avant de découvrir un terrain d'application privilégié de l'ANOVA : la méthode du plan d'expérience.

1 Les différentes analyses de variance

Différents types d'analyses de variance existent. La figure 5.1 présente le type d'analyse selon la nature des variables dépendantes et indépendantes. Nous commençons par détailler les principes de l'analyse univariée de la variance avant de passer à l'analyse multivariée puis à l'analyse de covariance.

Figure 5.1

Type d'analyse de variance selon la nature des variables.



Source : adapté de Malhotra (2007).

1.1 Les principes de l'analyse de variance

L'analyse de variance entre dans le cadre général du **modèle linéaire**, où une variable quantitative (ou plusieurs) est expliquée par une variable qualitative (ou plusieurs). L'objectif essentiel est de **comparer les moyennes empiriques de la variable quantitative observées pour les variables qualitatives** (facteurs) ou quantitatives découpées en classes (niveaux). À titre d'exemple, on étudiera la satisfaction du client à l'égard d'un produit selon les différentes caractéristiques de ce produit (parfum, texture, etc.).

Il s'agit de savoir si un facteur, ou une combinaison de facteurs (interaction), a un effet sur la variable quantitative expliquée. Par exemple, il s'agira de déterminer les caractéristiques optimales d'un paquet de céréales pour un petit déjeuner destiné aux enfants. Des indicateurs statistiques permettent de tester la significativité de cette combinaison linéaire.

1.2 L'analyse univariée de la variance : ANOVA à un facteur

L'analyse de variance sert à tester l'**hypothèse d'égalité des moyennes.** Cette technique est une extension du test *t* pour deux échantillons indépendants. Elle permet de traiter les différences de moyennes d'une variable dépendante quantitative Y lorsque la variable indépendante a plus de deux modalités. Ce type d'ANOVA permet de savoir si au moins une des moyennes diffère des autres. Ainsi, le salaire (variable quantitative) peut-il être expliqué par le diplôme (variable qualitative).

L'hypothèse nulle est vérifiée par le test *F* sous SPSS. Afin d'identifier les moyennes qui diffèrent, on peut comparer les moyennes avec les contrastes ou avec les tests post hoc.

Deux conditions sont nécessaires pour que les conclusions d'une ANOVA soient valides : l'**homogénéité** de la **variance intragroupe** et la **normalité des données**. Le **test de Levene** (> 0,05) est utilisé pour accepter l'hypothèse d'homogénéité de la variance intragroupe. Il faut, par ailleurs, vérifier la présence éventuelle de mesures aberrantes par le contrôle de la distribution des résidus à l'aide d'un graphique, les erreurs de saisie pouvant générer des hétérogénéités.

Si les données ne sont pas appropriées à une ANOVA (hétérogénéité des variances ou données fortement asymétriques), on doit alors utiliser des **tests non paramétriques** qui ne supposent ni homogénéité de la variance, ni une distribution normale, par exemple le test de Kruskall-Wallis.

Focus 5.1 Les tests post hoc et de comparaisons multiples

Lorsqu'on a déterminé qu'il existe des différences parmi les moyennes, les tests d'intervalle post hoc et de comparaisons multiples par paires déterminent les moyennes qui diffèrent. Ces tests servent à connaître, parmi plusieurs niveaux de modalités, ceux qui sont significativement différents des autres. Ils sont utilisés après que l'analyse de variance a été effectuée, si un facteur est significatif, et ils ne concernent que les facteurs ayant plus de deux niveaux.

Les tests post hoc les plus courants sont ceux de **Duncan**, de **Tukey**, de **Scheffé** et de **Bonferroni**. Le test de Duncan compare des moyennes deux à deux et suit un ordre pas à pas. Il utilise la statistique d'intervalle studentisé. Le test de Bonferroni, fondé sur la statistique *t* de Student, ajuste le niveau de signification observé en fonction du nombre de comparaisons multiples qui sont effectuées. Pour comparer un grand nombre de paires de moyennes, le test de Tukey est plus efficace que celui de Bonferroni. Le niveau de signification du test de Scheffé permet toutes les combinaisons linéaires possibles des moyennes de groupes à tester. Ce test est donc souvent plus strict que les autres; une plus grande différence de moyenne est nécessaire pour qu'il soit significatif.

SPSS

Étude du format du challenge avec une ANOVA à un facteur

De plus en plus d'entreprises organisent des challenges ayant un format de compétition mixte, c'est-à-dire comptant à la fois des objectifs individuels et des objectifs collectifs. Nous cherchons à connaître l'attitude des vendeurs à l'égard de ce nouveau format de compétition : le préfèrent-ils aux deux autres formats ?

Ouvrez le fichier « challenge » [1], allez dans le menu **Analyse > Comparer les moyennes > Anova à 1 facteur**.

La boîte de dialogue apparaît (voir figure 5.2), transférez les variables en les sélectionnant une à une puis en cliquant sur les flèches. La variable dépendante à tester est l'attitude à l'égard du challenge (ATTITUDECHALLENGE) et la variable indépendante est placée dans le champ **Facteur**.

Avant de lancer l'ANOVA à un facteur, nous vérifions l'homogénéité des moyennes.

Cliquez sur **Option**. Dans la boîte de dialogue qui apparaît (voir figure 5.3), cliquez sur **Test d'homogénéité**.

Cliquez sur **Poursuivre** pour revenir à la boîte de dialogue **MLG Univarié** puis sur **OK**.

1. Vous trouverez ce fichier à l'adresse : http://www.pearsoneducation.fr.

Figure 5.2

Commande d'une ANOVA à 1 facteur.

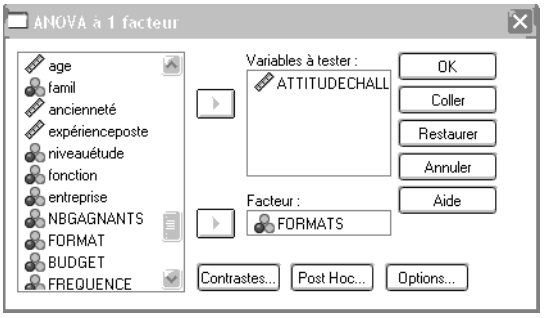
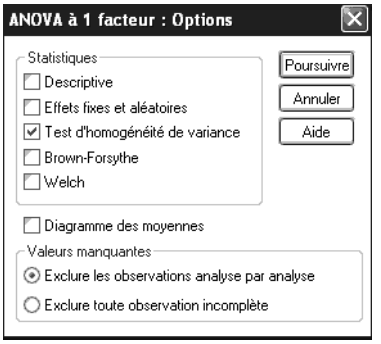


Figure 5.3

Test d'homogénéité pour ANOVA à 1 facteur.



Le test de Levene (voir figure 5.4) est significatif (0,52 > 0,05), l'hypothèse d'homogénéité des échantillons est donc acceptée. Nous pouvons procéder à l'analyse des résultats de l'ANOVA.

Figure 5.4

Interprétation du test d'homogénéité.

Test d'homogénéité des variances

ATTITUDECHALLENGE

Statistique de Levene	ddl1	ddl2	Signification
,653	2	698	,521

ANOVA

ATTITUDECHALLENGE

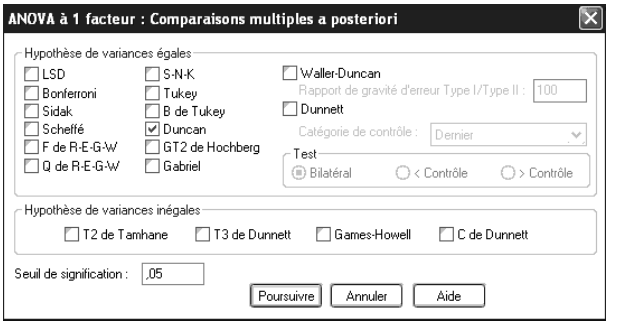
	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Intergroupes	11,676	2	5,838	5,254	,005
Intragroupes	775,631	698	1,111		
Total	787,307	700			

Plus la valeur de p est petite, plus la preuve est forte contre l'hypothèse nulle. Ici, les moyennes sont très différentes ($F = 5{,}25$; $p = 0{,}005$). L'hypothèse nulle est rejetée, le format a bien un effet sur l'attitude des commerciaux à l'égard du challenge mais, à ce stade, nous ne savons pas quel est leur type de format préféré. Pour cela, il faut réaliser un test de comparaisons multiples, aussi appelé test post hoc.

Cliquez sur **Post Hoc**. Dans la boîte de dialogue qui apparaît (voir figure 5.5), cliquez sur le test de votre choix. Nous sélectionnons ici le test de Duncan, souvent employé pour des tests de comparaisons de plusieurs moyennes. Dans notre cas, il s'agira des formats mixte, individuel ou en équipe.

Figure 5.5

Demande de test de comparaisons multiples *a posteriori* pour ANOVA à 1 facteur.



Cliquez sur **Poursuivre** pour revenir à la boîte de dialogue **ANOVA à un facteur** (voir figure 5.5) puis sur **OK**.

Le test de Duncan montre (voir figure 1.6) que le format de compétition en équipe est supérieur aux autres. L'attitude moyenne à son égard est de 2,97, significativement plus élevée que celle des formats individuel (2,69) ou mixte (2,64).

En d'autres termes, les vendeurs préfèrent un format de compétition en équipe plutôt qu'individuel ou mixte (avec des objectifs à la fois individuels et collectifs).

Figure 5.6

Résultats ANOVA à 1 facteur.

Tests post hoc

Sous-ensembles homogènes

ATTITUDECHALLENGE

Duncan[a,b]

FORMATS	N	Sous-ensemble pour alpha = .05	
		1	2
mixte	315	2,6413	
individuel	236	2,6935	
équipe	150		2,9733
Signification		,609	1,000

Les moyennes des groupes des sous-ensembles homogènes sont affichées.

a. Utilise la taille d'échantillon de la moyenne harmonique = 213,090.

b. Les effectifs des groupes ne sont pas égaux. La moyenne harmonique des effectifs des groupes est utilisée. Les niveaux des erreurs de type I ne sont pas garantis.

1.3 L'ANALYSE DE VARIANCE À X FACTEURS

L'ANOVA à plusieurs facteurs consiste à tester l'hypothèse d'égalité des moyennes d'une variable quantitative selon plusieurs variables qualitatives. Par exemple, on pourra tester les variations de salaire selon le diplôme et le sexe de l'employé. Le test de significativité est un **test *F***. Il s'appuie sur la **décomposition de la variance** qui comprend : les **effets principaux**, les **effets d'interaction** et un terme résiduel. La notion d'interaction correspond au fait que l'effet d'une variable explicative sur la variable à expliquer n'est pas identique selon le niveau de l'autre variable explicative (Evrard *et al.*, 2003). L'interaction peut être ordinale, l'ordre des effets liés au premier facteur respecte alors celui des niveaux du second facteur. Dans le cas d'une interaction non ordinale, nous avons en revanche une modification dans l'ordre des effets. Une interaction non ordinale peut être non croisée ou croisée. Cette dernière interaction est la plus forte de toutes.

L'existence d'une interaction se détecte par l'observation des courbes : leur parallélisme indique une absence d'interaction (l'effet conjoint des facteurs combinés est égal à la somme de leurs effets principaux individuels) alors que leur croisement montre que l'effet relatif des niveaux du premier facteur varie en fonction des niveaux de l'autre.

SPSS

Étude de l'impact de trois modalités des challenges avec une ANOVA

Lors des challenges, les vendeurs changent souvent leur manière de vendre ; ils seraient plus agressifs afin de gagner et moins attentifs aux attentes de leurs clients. Nous cherchons à savoir comment ils réagissent, quelle est leur orientation client (OC) selon trois caractéristiques des challenges : la fréquence de ceux-ci, le nombre de gagnants et le format de compétition du challenge.

Ouvrez le fichier exemple « challenge », disponible sur le site de l'ouvrage [1].

Allez dans le menu **Analyse > Comparer les moyennes > Anova à un facteur**. Dans la boîte de dialogue qui apparaît (voir figure 5.7), transférez les variables en les sélectionnant une à une puis en cliquant sur les flèches. La variable dépendante est l'orientation client lors du challenge (OCCHALLENGE) et les variables indépendantes ou explicatives sont placées dans le champ **Facteur(s) fixe(s)**. Ici, les modalités des challenges sont : fréquence, format de compétition, nombre de gagnants (FREQUENCE, NBGAGNANTS, FORMAT).

Quelques remarques sur les boutons de cette boîte de dialogue :

Modèle. Ce bouton permet de préciser le type de modèle à analyser et le type d'erreur. Par défaut, sont cochées l'option plan complet, où tous les effets sont calculés, et l'erreur de type III, qui permet de tester des modèles équilibrés ou non (un modèle est déséquilibré lorsque les cellules ne contiennent pas le même nombre d'observations). Il faut cliquer sur le bouton Autre, faire passer les variables dans la partie Modèle et choisir les effets (principaux, d'interaction, d'ordre 2, etc.) pour en analyser seulement certains.

Contrastes. Sert à tester les différences entre les niveaux des facteurs.

Diagrammes. Ce bouton permet de comparer avec des graphiques les moyennes de la variable dépendante selon le niveau de chaque facteur.

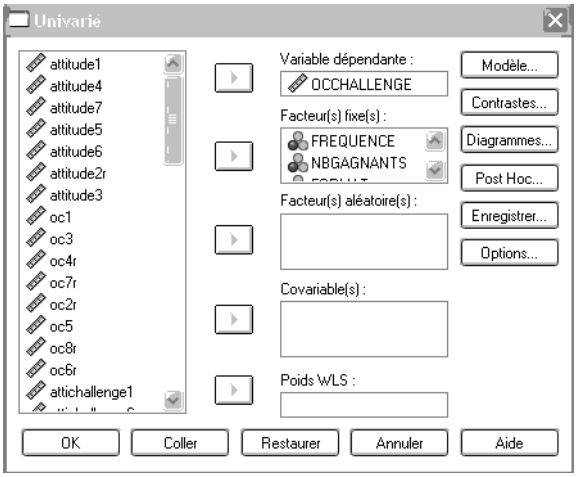
Post Hoc. Permet d'identifier, parmi plusieurs moyennes, celles qui diffèrent (voir focus 5.1).

Enregistrer. Permet de sauvegarder les valeurs prédites avec le modèle, les résidus et les autres mesures comme nouvelles variables dans l'éditeur de données.

1. Vous trouverez ce fichier à l'adresse : http://www.pearsoneducation.fr.

Figure 5.7

Commande d'une ANOVA à 3 facteurs.



Options. Pour obtenir diverses statistiques, par exemple, le test d'homogénéité des variances. Il permet aussi de spécifier le seuil de significativité (fixé par défaut à 0,05) pour l'étude des comparaisons de moyennes et le calcul d'intervalles de confiance.

Avant de lancer l'ANOVA, nous devons vérifier l'homogénéité des moyennes.

Cliquez sur **Option**. Dans la boîte de dialogue qui apparaît (voir figure 5.8), cliquez sur **Tests d'homogénéité**.

Figure 5.8

Tests d'homogénéité pour ANOVA.

Cliquez sur **Poursuivre** pour revenir à la boîte de dialogue **MLG Univarié** puis sur **OK**.

Le test de Levene (voir figure 5.9) est significatif (0,18 > 0,05), l'hypothèse d'homogénéité des échantillons est donc acceptée. Nous pouvons procéder à l'analyse des résultats de l'ANOVA (voir figure 5.10).

Le modèle explique 5 % de l'orientation client du vendeur pendant un challenge. Le nombre de gagnants (F = 6,70 ; p = 0,01) et l'interaction de fréquence/format (F = 4,34 ; p = 0 03) ont un impact significatif sur l'orientation client.

Figure 5.9

Interprétation du test d'homogénéité.

Test d'égalité des variances des erreurs de Levene[a]

Variable dépendante : OCCHALLENGE

F	ddl1	ddl2	Signification
1,459	7	380	,180

Teste l'hypothèse nulle que la variance des erreurs de la variable dépendante est égale sur les différents groupes.

a. Plan : Ordonnée à l'origine+FREQUENCE+NBGAGNANTS+FORMAT+FREQUENCE * NBGAGNANTS+FREQUENCE * FORMAT+NBGAGNANTS * FORMAT+FREQUENCE * NBGAGNANTS * FORMAT

Figure 5.10

Résultats de l'ANOVA à 3 facteurs.

Tests des effets intersujets

Variable dépendante: OC

Source	Somme des carrés de type III	ddl	Moyenne des carrés	F	Signification
Modèle corrigé	17,890[a]	7	2,556	3,244	,002
Constante	1589,924	1	1589,924	2017,942	,000
FREQUENCE	2,304	1	2,304	2,925	,088
NBGAGNANTS	5,281	1	5,281	6,703	,010
FORMAT	,087	1	,087	,110	,740
FREQUENCE * NBGAGNANTS	,069	1	,069	,087	,768
FREQUENCE * FORMAT	3,420	1	3,420	4,341	,038
NBGAGNANTS * FORMAT	,032	1	,032	,041	,840
FREQUENCE * NBGAGNANTS * FORMAT	2,120	1	2,120	2,691	,102
Erreur	299,400	380	,788		
Total	2329,833	388			
Total corrigé	317,290	387			

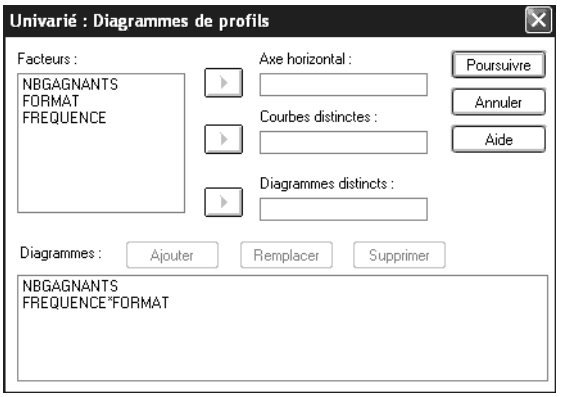
a. R deux = ,056 (R deux ajusté = ,039)

Cependant, à ce stade, nous ne savons pas lequel des deux niveaux affecte le moins l'orientation client du vendeur. Les facteurs manipulés ayant chacun deux niveaux, il est possible de visualiser directement leurs effets à l'aide de graphiques.

Allez dans le menu **Analyse > Modèle Linéaire Général > Univarié**. Dans la boîte de dialogue qui apparaît, cliquez sur le bouton **Diagrammes**. Faites passer les facteurs significatifs dans Axe horizontal et Courbes distinctes (effets d'interaction) puis cliquez sur **Ajouter** (voir figure 5.11).

Figure 5.11

Obtention de graphique afin de visualiser les effets des facteurs significatifs.

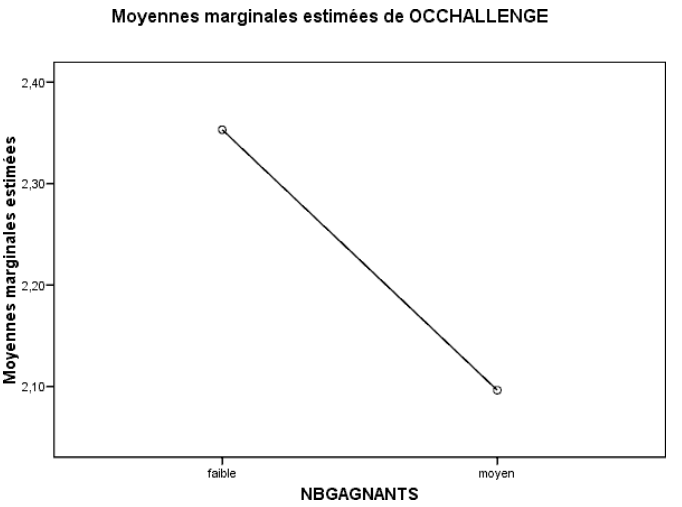


Cliquez sur **Poursuivre** pour revenir à la boîte de dialogue **MLG Univarié** puis sur **OK**.

Le graphique (voir figure 5.12) atteste que lorsque le nombre de gagnants au challenge est faible, l'orientation client des commerciaux baisse moins que lorsque le challenge fait un nombre moyen de gagnants. Le challenge qui préserve mieux l'orientation client fait donc peu de vainqueurs.

Figure 5.12

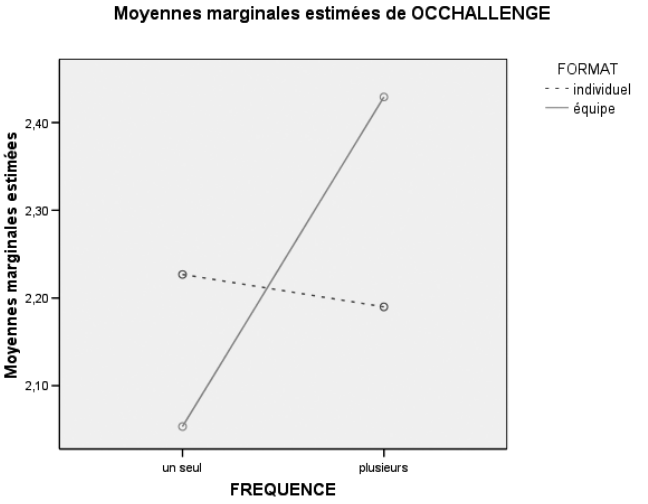
Graphique d'un facteur ayant un effet principal significatif.



Nous constatons (voir figure 5.13) que le challenge en équipe avec une fréquence faible (un seul challenge organise à la fois) maximise l'orientation client du vendeur.

Figure 5.13

Graphique de facteurs dont l'effet d'interaction est significatif.



1.4 L'ANALYSE MULTIVARIÉE DE LA VARIANCE : MANOVA

L'analyse multivariée de la variance est une extension de l'ANOVA pour mesurer les différences de moyenne de deux variables dépendantes quantitatives (ou plus) en fonction de plusieurs variables qualitatives. Ce que la MANOVA apporte en plus de l'analyse de variance simple concerne la corrélation des variables à expliquer, décomposée en **intra et intergroupe**. Ces interactions apparaissent lorsque les effets d'un facteur donné sur les variables expliquées dépendent des modalités des autres facteurs.

Un des apports majeurs de l'analyse multivariée de la variance est la mise en évidence, parmi un ensemble de variables quantitatives, de celles dont la valeur est la plus affectée par les variations d'un ensemble de variables quantitatives ou qualitatives. Cela tient compte des **intercorrélations** entre variables à expliquer ; nous pouvons ainsi déceler les variables qui contribuent le plus à la formation de la combinaison linéaire pour les soumettre ensuite à une analyse de variance univariée (Evrard *et al.*, 2003).

En règle générale, les modalités de la variable indépendante sont présumées fixes (modèle à effet fixe).

Plusieurs conditions sont à valider lors de l'utilisation d'une MANOVA (Hair *et al.*, 2006) :

- **Seuil d'observations par cellule** de 20 ou au minimum supérieur au nombre de variables dépendantes.
- **Indépendance des observations.**
- **Égalité des matrices de variance-covariance** entre les groupes. La violation de cette hypothèse, vérifiée par **le test de Box**, n'a toutefois qu'un impact limité si les groupes sont de taille à peu près identique. Il est généralement recommandé d'avoir un rapport, entre la taille du groupe le plus important et celle du plus grand groupe, inférieur à 1,5.
- **Multinormalité des distributions des moyennes** pour chaque variable dépendante. La MANOVA est cependant robuste en cas de violation de cette hypothèse si la taille des groupes est importante.
- **Linéarité des variables dépendantes.**
- **La significativité des différences de moyennes s'appuie sur le test *F*,** complété d'autres statistiques : la trace de Hotelling, la plus grande racine de Roy, le lambda de Wilks et le critère de Pillai-Bartlett pour chaque variable explicative. Si **le critère de Pillai** est le plus robuste à la violation de certaines conditions d'utilisation de la MANOVA, il est conseillé de comparer cette statistique aux autres indicateurs.

L'interprétation d'une MANOVA se fait en deux temps. À la première étape, l'examen du critère de Pillai indique la significativité des variables explicatives. À la seconde, il faut déceler, parmi les variables expliquées, celles qui sont affectées par la variable indépendante. Les résultats se lisent alors, comme pour une ANOVA, sur les courbes des diagrammes.

1.5 L'analyse de covariance : ANCOVA et MANCOVA

L'analyse de covariance (ANCOVA) combine les **techniques de l'analyse de variance et de la régression**. La MANCOVA est une extension des principes de l'ANCOVA à l'analyse multivariée, c'est-à-dire sur plusieurs variables dépendantes.

Ces méthodes sont recommandées pour éliminer des erreurs systématiques hors du contrôle du chercheur et pouvant biaiser les résultats. L'ajout d'une covariable peut éliminer une source potentielle de variance qui aurait appartenu à l'erreur expérimentale si elle avait été ignorée (Nunnally et Bernstein, 1994). Cependant, trop de covariables réduit l'efficience statistique des procédures. Une règle de base est d'avoir un nombre de covariables inférieur à : (0,1 × la taille de l'échantillon) – (nombre de groupes – 1) [Hair *et al.*, 2006].

Afin d'améliorer l'analyse de covariance, il faut essayer de minimiser le nombre de covariables tout en s'assurant que les plus importantes ne sont pas éliminées. Une covariable est pertinente si elle est corrélée à la variable dépendante et non corrélée à la (ou aux)

variable(s) indépendante(s). Une autre fonction de l'ANCOVA réside aussi dans la prise en compte des différences de réponses dues aux caractéristiques des répondants.

L'interprétation d'une ANCOVA et d'une MANCOVA se fait en deux étapes. En premier lieu, il faut considérer l'effet de la (ou des) covariable(s). Ce résultat se lit comme celui d'une régression. Ensuite, il faut interpréter les résultats des facteurs explicatifs (variables qualitatives).

SPSS

Réalisation d'une ANCOVA

Nous cherchons à expliquer l'attitude du vendeur à l'égard du challenge. Pour cela, nous avons réalisé une ANOVA ayant pour facteurs la fréquence, le format et le nombre de gagnants. Pour améliorer la précision de ces résultats, nous ajoutons des variables qui pourraient expliquer les variations d'attitude des commerciaux. À ce titre, deux nouveaux éléments sont pris en compte : l'*entreprise*, c'est-à-dire l'établissement bancaire où travaille chaque vendeur, dont nous pouvons supposer qu'il influe sur les résultats et, la *fonction* du vendeur. En effet, l'échantillon étant composé de vendeurs issus d'entreprises et de fonctions différentes, il est possible que ces variables *entreprise* et *fonction* soit à l'origine de variations dans les réponses. L'intégration de ces covariables améliore la précision de l'analyse et permet de redresser les biais dus au fait que les répondants ont différentes responsabilités. En conséquence, l'entreprise et la fonction du commercial sont ajoutées aux variables explicatives pour toutes les variables expliquées afin de les contrôler.

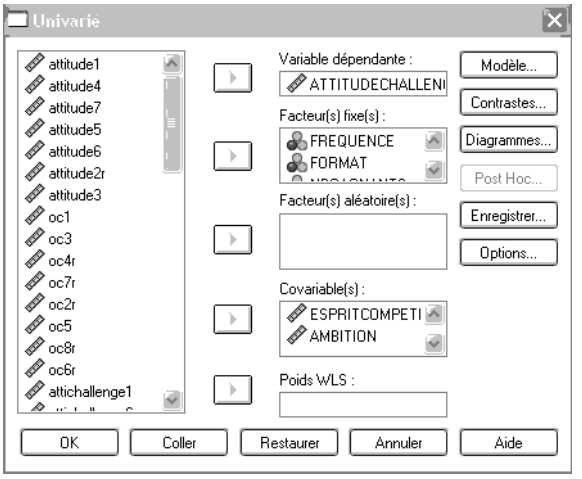
Par ailleurs, deux caractéristiques individuelles des vendeurs sont des variables explicatives importantes du processus de motivation : l'esprit de compétition et l'ambition du vendeur. Nous testerons donc les effets de ces covariables sur l'attitude du vendeur à l'égard du challenge.

Ouvrez le fichier exemple « challenge », disponible sur le site de l'ouvrage [1].

Allez dans le menu **Analyse > Modèle Linéaire Général > Univarié**. Dans la boîte de dialogue qui apparaît (voir figure 5.14), transférez les variables en les sélectionnant une à une puis en cliquant sur les flèches. La variable dépendante est l'attitude à l'égard du challenge (ATTITUDECHALLENGE). Les variables indépendantes sont placées dans les champs Facteur(s) fixe(s) et Covariable(s). Les facteurs fixes sont : les modalités des challenges (FREQUENCE, FORMAT, NBGAGNANTS) ; les covariables sont : l'entreprise, la fonction du vendeur (entreprise, fonction), l'esprit de compétition et l'ambition.

Figure 5.14

Commande d'une ANCOVA.



1. Vous trouverez ce fichier à l'adresse : http://www.pearsoneducation.fr.

Le test de Levene (voir figure 5.15) est significatif (0,22), l'hypothèse d'homogénéité des échantillons est donc acceptée. Nous pouvons procéder à l'analyse des résultats de l'ANCOVA (voir figure 5.16).

Les caractéristiques des challenges ainsi que les covariables expliquent 55 % de l'attitude du vendeur à l'égard du challenge (voir figure 5.16).

Figure 5.15

Interprétation du test d'homogénéité.

Test d'égalité des variances des erreurs de Levene[a]

Variable dépendante: ATTITUDECHALLENGE

F	ddl1	ddl2	Signification
1,357	7	331	,223

Teste l'hypothèse nulle que la variance des erreurs de la variable dépendante est égale sur les différents groupes.

a. Plan : Ordonnée à l'origine+entreprise+fonction+ESPRITCOMPETITION+AMBITION+FREQUENCE+FORMAT+NBGAGNANTS+FREQUENCE * FORMAT+FREQUENCE * NBGAGNANTS+FORMAT * NBGAGNANTS+FREQUENCE * FORMAT * NBGAGNANTS

Figure 5.16

Résultats de l'ANCOVA à cinq variables explicatives.

Tests des effets inter-sujets

Variable dépendante: ATTITUDECHALLENGE

Source	Somme des carrés de type III	ddl	Moyenne des carrés	F	Signification
Modèle corrigé	208,039[a]	11	18,913	37,606	,000
Constante	1,285	1	1,285	2,554	,111
entreprise	8,942	1	8,942	17,780	,000
fonction	,054	1	,054	,108	,742
ESPRITCOMPETITION	101,352	1	101,352	201,530	,000
AMBITION	,035	1	,035	,070	,791
FREQUENCE	1,048	1	1,048	2,084	,150
FORMAT	4,118	1	4,118	8,187	,004
NBGAGNANTS	,003	1	,003	,006	,937
FREQUENCE * FORMAT	,855	1	,855	1,700	,193
FREQUENCE * NBGAGNANTS	,088	1	,088	,174	,677
FORMAT * NBGAGNANTS	,620	1	,620	1,234	,268
FREQUENCE * FORMAT * NBGAGNANTS	,516	1	,516	1,027	,312
Erreur	164,453	327	,503		
Total	3113,778	339			
Total corrigé	372,492	338			

a. R deux = ,559 (R deux ajusté = ,544)

Les résultats de l'ANCOVA montrent que les covariables *entreprise* et *esprit de compétition* ont un effet significatif sur l'attitude du vendeur à l'égard du challenge. Comme nous l'avions postulé, l'esprit de compétition du vendeur a un effet significatif, cependant, à ce stade, nous ne savons pas s'il est positif.

Pour le découvrir, retournez dans la boîte de dialogue : **Analyse > Modèle Linéaire Général > Univarié** et cliquez sur **Options** puis sur **Estimations des paramètres** (voir figure 5.17). Les résultats apparaissent alors pour l'ensemble des variables et des covariables.

Les résultats de cette commande se lisent à la figure 5.18.

Le tableau d'estimations des paramètres montre que, conformément à ce qui apparaît à la figure 5.18, l'entreprise et l'esprit de compétition ont un effet significatif sur l'attitude du vendeur à l'égard du challenge. L'esprit de compétition a un effet positif sur cette attitude ($\beta = 0{,}77$; $t = 14{,}19$).

Comme pour une ANOVA, la lecture des effets des variables qualitatives se fait à l'aide d'un graphique. La figure 5.19 atteste que le format a un impact significatif sur l'attitude du vendeur. Pour commander un diagramme pour ce facteur, allez dans le menu **Analyse > Modèle Linéaire Général > Univarié** puis cliquez sur le bouton **Diagrammes**.

Figure 5.17

Commande des résultats des covariables.

Figure 5.18

Interprétation des covariables.

Estimations des paramètres

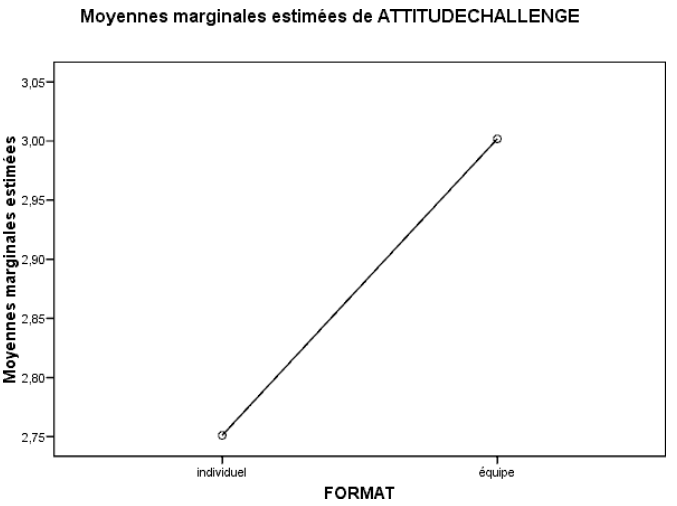
Variable dépendante : ATTITUDECHALLENGE

Paramètre	B	Erreur standard	t	Signification	Intervalle de confiance à 95%	
					Borne inférieure	Limite supérieure
Constante	,158	,210	,752	,452	-,255	,570
entreprise	,156	,037	4,217	,000	,083	,228
fonction	-,016	,047	-,329	,742	-,109	,078
ESPRITCOMPETITION	,777	,055	14,196	,000	,669	,884
AMBITION	,016	,059	,265	,791	-,100	,131

Dans la boîte de dialogue Diagrammes de profils, faites passer le format dans Axe horizontal puis cliquez sur **Ajouter**. Le graphique visible figure 5.19 apparaît.

Figure 5.19

Interprétation des variables explicatives de l'ANCOVA.



Le graphique montre que le format qui maximise l'attitude du vendeur à l'égard du challenge est le format de compétition en équipe.

Cet exemple dévoile comment une ANCOVA prend en compte des variables explicatives à la fois qualitatives et quantitatives dans un même traitement. La MANCOVA va plus loin puisqu'elle traite en même temps plusieurs variables explicatives de différentes natures ainsi que plusieurs variables dépendantes quantitatives. Par exemple, elle permet d'étudier les effets des caractéristiques de challenges et de l'esprit de compétition des vendeurs sur leur attitude à l'égard des challenges ainsi que sur leurs comportements à l'égard de la clientèle (orientation client). De fait, il est possible d'étudier le challenge optimal en termes de satisfaction vendeur et de satisfaction client.

Concernant les covariables, les résultats de la MANCOVA et ceux de l'ANCOVA se lisent de façon similaire. D'autres statistiques, telles que la trace de Hotelling, la plus grande racine de Roy, le lambda de Wilks et le critère de Pillai-Bartlett, servent à interpréter, pour chaque variable explicative, leurs effets sur les variables dépendantes.

La méthode du plan d'expérience, que nous allons explorer, est un terrain d'application privilégié de l'analyse de variance.

2 La méthode du plan d'expérience

L'expérimentation fait partie de notre quotidien. Nous cherchons souvent à connaître l'effet de facteurs sur divers résultats. La plupart du temps, cela se fait de manière informelle, par exemple, en se posant les questions : Est-ce qu'en partant trente minutes plus tôt au travail, j'aurai plus de chance de trouver une place de parking ? *Quid* de vingt minutes ou de dix minutes ?

La méthode du plan d'expérience a été mise au point dans les années 1920, par Ronald A. Fisher, dans le cadre d'études agronomiques. Son utilisation s'est développée en sciences sociales et en marketing depuis une trentaine d'années.

Après avoir présenté l'expérimentation, nous aborderons des exemples de plans d'expérience permettant d'en comprendre l'utilité.

2.1 La méthode expérimentale en sciences de gestion

L'expérimentation est une opération où l'on cherche à vérifier des relations de cause à effet par manipulation de facteurs. Il s'agit de manipuler une variable indépendante (ou plusieurs) et d'en mesurer l'effet sur une variable dépendante (ou plusieurs), cela en contrôlant les variables externes susceptibles d'influer sur les résultats.

L'avantage majeur de cette méthode est donc l'isolement de ce qui est dû à la variable déterminante examinée. Son inconvénient principal réside dans la validité externe limitée de l'expérience. En outre, des variables externes, ou biais, peuvent brouiller les mesures des variables dépendantes et affaiblir la validité des résultats.

L'expérimentation peut s'appuyer sur la méthode dite des scénarios ou des vignettes qui est ici détaillée.

Focus 5.2 La méthode des scénarios ou des vignettes

La méthode dite des **scénarios** ou des **vignettes** est issue des **techniques projectives** utilisées en psychologie et dont l'objectif est de s'intéresser aux attitudes et aux comportements des acteurs. Le scénario est une brève histoire qui, si elle est soigneusement élaborée, simule de vraies expériences de la vie. Les individus, mis dans une situation hypothétique, doivent répondre « comme si » ils se trouvaient réellement dans ces situations. La mise en situation présente l'avantage de rendre l'expérience plus réaliste et, par là, de mieux impliquer les répondants dans la création de sens.

Une description verbale, un texte descriptif écrit, une photo, un dessin ou un prototype peuvent servir à présenter **les *stimuli***. Il faut ici veiller à ce que les scénarios soient crédibles et qu'aucun ne soit manifestement trop attractif ou, au contraire, répulsif. Les *stimuli* doivent de fait avoir une apparence similaire afin que les préférences des individus soient bien le fruit des attributs testés.

Les biais de l'expérimentation

Afin de renforcer la validité de l'expérience, il faut connaître les principaux biais expérimentaux :

- **Histoire.** Un événement porteur d'effets se produit entre deux mesures, par exemple, la crise de la vache folle a changé la perception de la qualité de la viande en France.
- **Maturation.** Des changements se produisent chez les individus au fil de l'expérience (fatigue, désintérêt).
- **Effet de test.** La situation d'expérience provoque par elle-même un biais. Par exemple, la réponse à une première question suscite une réflexion qui va modifier les réponses aux questions suivantes.
- **Effet de l'instrument.** L'application de l'instrument de mesure fausse le résultat.
- **Mortalité.** Personnes de l'échantillon initial qui ne veulent plus répondre dans le cas où l'expérimentation s'effectue en plusieurs étapes (étude longitudinale).

Typologie des plans d'expérience

Voici une typologie simplifiée des plans d'expérience : **préexpérimentaux** (étude de cas unique, prétest/post-test sur un seul groupe, groupe statique), **expérimentaux** (prétest/post-test ou seulement post-test avec groupe de contrôle) et **modèles statistiques** (bloc aléatoire, carré latin, plan factoriel) [Malhotra *et al.*, 2007].

- **Étude de cas unique.** Un seul groupe d'individus (ou d'autres entités) est exposé à une variable et on ne prend qu'une seule mesure de la variable dépendante. La sélection des individus est réalisée de manière arbitraire. Le problème est que cette étude ne permet pas d'obtenir le niveau de la variable expliquée s'il n'y avait pas eu d'exposition.
- **Prétest/post-test et groupe de contrôle.** Lorsque l'expérience inclut un prétest, les individus évaluent deux fois la variable dépendante : avant et après le test. L'effet d'expérimentation est alors calculé (mesure post – mesure prétraitement).
- **Groupe statique.** Le plan expérimental comprend deux groupes : un exposé à la variable indépendante et l'autre non (groupe de contrôle). L'effet de l'expérimentation est mesuré (mesure du groupe de contrôle – mesure du groupe expérimental).
- **Modèles statistiques.** Ils permettent de mesurer les effets de plus d'une variable indépendante simultanément et de contrôler statistiquement des variables externes précises. On

distingue le bloc aléatoire et le plan factoriel. Le bloc aléatoire regroupe les individus en fonction d'une seule variable externe majeure (par exemple, le type de client) susceptible d'influer sur la variable dépendante. Les répondants de chaque bloc sont affectés aléatoirement aux groupes de traitement. Le plan factoriel autorise l'étude de deux variables externes non interactives ou plus et d'une ou plusieurs variables indépendantes.

Nous allons aborder maintenant le plan factoriel et ses différentes versions.

2.2 Le plan factoriel

Le plan factoriel sert à mesurer les effets de plusieurs variables indépendantes ayant plusieurs niveaux différents. Il permet l'étude à la fois des effets principaux et des effets d'interaction de ces niveaux. Par exemple, on pourra examiner l'effet du type de carburant et du type de conduite sur la consommation de carburant.

La notion d'interaction correspond au fait que l'effet d'une variable explicative sur une variable à expliquer est changeant selon le niveau de l'autre variable explicative. Il y a interaction quand l'effet simultané de plusieurs variables diffère de leurs effets séparés cumulés. Par exemple, un individu peut préférer sortir au cinéma (plutôt que d'aller au théâtre, au concert ou encore chez des amis) et l'été peut être sa saison favorite pour sortir (plutôt que les trois autres saisons), mais il peut ne pas préférer aller au cinéma l'été.

D'un point de vue statistique, un **plan factoriel** est l'agrégation de variables indépendantes : ensemble de niveaux de chaque variable indépendante et combinaisons de ces niveaux sélectionnés pour l'expérimentation. Le dispositif expérimental doit permettre de répondre aux trois questions suivantes :

1. Quels sont les facteurs fondamentaux sélectionnés ?
2. Comment les niveaux de ces facteurs varient-ils ?
3. Comment ces niveaux doivent-ils être combinés ?

Par exemple, pour une étude de packaging de céréales pour petit déjeuner, on se demandera :

1. Quels facteurs sélectionne-t-on ? La forme du paquet, les codes couleur, le style, le matériau utilisé ?

   Ensuite, si l'on choisit d'étudier la taille du paquet et le style :

2. Combien de niveaux choisit-on d'examiner ? Pour la taille du paquet : grand, moyen, petit ? Pour le style : sport, santé, régime ?
3. Croise-t-on tous ces niveaux ou seulement les combinaisons les plus pertinentes ?

La sélection des facteurs et des niveaux étudiés revient au chargé d'étude qui estime les variables les mieux à même d'expliquer la variable dépendante. Le choix des combinaisons à tester détermine ensuite celui du plan factoriel utilisé (complet ou fractionné). Souvent, l'étude d'un grand nombre de facteurs et de niveaux conduit à employer un plan fractionné.

2.3 Plan factoriel complet ou fractionné ?

Le **plan factoriel complet** permet de tester tous les niveaux de chacun des facteurs sur chacun des niveaux des autres facteurs. Dans notre exemple de packaging de céréales pour petit déjeuner, si l'on sélectionne trois niveaux de taille du paquet (grand, moyen et petit), deux niveaux de messages (sport et santé), un plan factoriel complet permettra de tester toutes les combinaisons possibles, c'est-à-dire 6 (3 × 2). Les influences de chaque facteur et de ses interactions avec tous les autres facteurs seront étudiées. S'il nécessite davantage d'expériences, ce plan est plus riche que le plan factoriel fractionné.

L'avantage du **plan factoriel fractionné**, constitué d'un sous-ensemble de combinaisons d'un plan complet, réside dans sa capacité à examiner un grand nombre de facteurs dont il serait difficile de tester tous les niveaux. Il permet ainsi de réaliser des économies substantielles d'expériences. Toujours avec notre exemple de packaging, un plan factoriel fractionné permettra d'examiner un plus grand nombre de modalités (taille du paquet, message, codes couleurs, style, etc.) et de niveaux (3, 4 niveaux pour chaque facteur examiné) tout en ne testant qu'un nombre restreint de paquets différents.

Malgré l'intérêt qu'il présente en termes d'économie d'expériences, ce type de plan compte des effets confondus. Ces effets gênent l'interprétation de certains effets principaux qui sont mélangés avec des interactions.

Focus 5.3 Les plans fractionnés en carrés latin et gréco-latin

Lorsque le chargé d'étude ne peut pas tester l'ensemble des attributs et de leurs niveaux parce qu'ils sont trop nombreux, il est fréquent d'avoir recours à un plan fractionné. Le **carré latin** et le **gréco-latin** (second carré latin sur un premier) sont souvent utilisés car ils permettent de faire des économies importantes d'expériences : 9 au lieu de 27 (3 × 3 × 3) combinaisons pour le carré latin ou 81 (3 × 3 × 3 × 3) pour le carré gréco-latin. Ces plans ou carrés ne croisent pas tous les facteurs. On peut par exemple tester l'influence de la fréquence de challenges de vente, du nombre de gagnants et du format de compétition en créant un niveau supplémentaire pour chacun de ces facteurs.

L'inconvénient principal des carrés latin et gréco-latin est donc l'obligation d'avoir, pour chaque facteur manipulé, le **même nombre de niveaux**. Autre problème important, ces plans ne permettent d'examiner que l'effet principal de chacun des facteurs et non leurs interactions.

Les deux exemples d'application suivants illustrent la réalisation d'une expérimentation avec un plan complet puis avec un plan fractionné.

Exemple

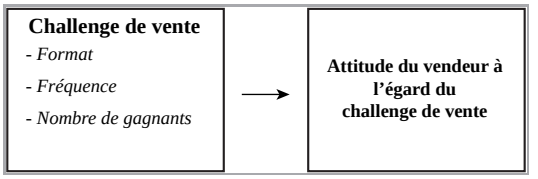
Étude des caractéristiques de challenges de vente avec un plan factoriel complet

À la suite d'entretiens avec des commerciaux, trois caractéristiques des challenges apparaissent très importantes dans la formation de leur attitude : le format de compétition, la fréquence des challenges et le nombre de gagnants. Nous cherchons à tester l'effet des trois facteurs ayant chacun deux niveaux (voir figure 5.20). Pour chacun des facteurs, les différents niveaux examinés seront les suivants :

- le format de compétition : individuel (1) ou en équipe (2) ;
- la fréquence : faible (1) ou élevée (2) ;
- le nombre de gagnants : faible (1) ou moyen (2).

Figure 5.20

Le plan testé.



L'étude de toutes les modalités des challenges et de leurs niveaux requiert l'utilisation d'un plan complet. Le protocole de l'expérimentation est par conséquent constitué de 2 × 2 × 2, soit 8 combinaisons de caractéristiques de challenges. Il faudra donc présenter aux individus huit challenges de vente différents.

Le **plan factoriel complet orthogonal** permet l'estimation de la moyenne des effets des facteurs sans craindre que les résultats subissent une distorsion par l'effet des autres facteurs. Toutes les interactions sont testées. L'**orthogonalité** est vérifiée en mettant en place ce protocole : (1) on remplace les valeurs 1, 2 dans la matrice plan par les valeurs –1, 1 respectivement ; (2) on additionne ensuite les valeurs correspondantes de chacune des colonnes ; (3), si la somme est égale à zéro, les colonnes sont orthogonales et les effets représentés par ces colonnes sont alors dits orthogonaux.

Tableau 5.1 : Plan factoriel complet

Scénarios	Format	Fréquence	Nb gagnants	Orthogonalité
n° 1	1	1	1	–3
n° 2	1	1	2	–1
n° 3	1	2	1	–1
n° 4	1	2	2	1
n° 5	2	1	1	–1
n° 6	2	1	2	1
n° 7	2	2	1	1
n° 8	2	2	2	3
				Σ = 0

Par exemple, le scénario n° 2 correspond ici à un challenge individuel, peu fréquent et faisant un nombre moyen de gagnants.

EXEMPLE

Étude des caractéristiques de challenges de vente avec un plan factoriel fractionné

Trois caractéristiques des challenges de vente sont maintenant étudiées avec, pour chacun de ces facteurs ou attributs, non plus deux mais trois niveaux :

- le format de compétition : individuel (A), en équipe (B) ou mixte (C) (objectifs individuels et collectifs) ;
- la fréquence : élevée, moyenne ou faible ;
- le nombre de gagnants : élevé, moyen ou faible.

Il faudrait normalement 3 × 3 × 3 = 27 combinaisons. Nous avons vu au focus 5.3 que le carré latin permettait de passer de 27 à seulement 9 combinaisons ou challenges testés. Voici comment se construit ce plan fractionné.

Tableau 5.2 : Carré latin

Fréquence	**Nb gagnants**			
	Élevé	**Moyen**	**Faible**	
Élevée	A – n° 1	B – n° 4	C – n° 7	ABC
Moyenne	C – n° 2	A – n° 5	B – n° 8	CAB
Faible	B – n° 3	C – n° 6	A – n° 9	BCA
	ACB	BAC	CBA	

Le format de compétition qui est la troisième variable manipulée est soit individuel (A), soit en équipe (B), soit mixte (C). Chacun de ses niveaux doit apparaître dans chaque ligne et chaque colonne.

Pour comprendre l'élaboration de ce plan, nous prenons l'exemple du format de compétition. Le format individuel (A) apparaît en colonne 1, ligne 1, en colonne 2, ligne 2 et en colonne 3, ligne 3 ; le format en équipe (B) figure en colonne 1, ligne 3, en colonne 2, ligne 1 et en colonne 3, ligne 2 ; le format mixte (C) figure en colonne 1, ligne 2, en colonne 2, ligne 3 et en colonne 3, ligne 1.

Chacun des niveaux de la fréquence et du nombre de gagnants est testé une fois avec tous les autres niveaux des deux autres facteurs. Il en est de même pour tous les niveaux du facteur nombre de gagnants.

Par exemple, le scénario n° 7 correspond ici à un challenge mixte, peu fréquent et faisant un faible nombre de gagnants.

Résumé

L'analyse de variance et de covariance sert à évaluer les écarts des valeurs moyennes d'une variable dépendante sous l'effet de variables indépendantes contrôlées et, cela, en tenant compte de l'influence de variables indépendantes non contrôlées. L'ANOVA permet d'effectuer un test sur les moyennes de deux populations ou plus. Le test *F* permet de vérifier l'hypothèse nulle qui suppose l'égalité des moyennes.

L'analyse multivariée de la variance implique l'examen simultané de plusieurs variables indépendantes qualitatives. Elle permet l'évaluation de l'interaction de ces variables. Le test *F* sert à vérifier la signification de l'effet global, des effets principaux et des interactions. Il y a interaction lorsque l'effet d'une variable indépendante sur une variable dépendante diffère en fonction des modalités ou niveaux d'une autre variable indépendante.

L'analyse de covariable fait référence, en plus de variable(s) indépendante(s) qualitative(s), au test de variable(s) indépendante(s) quantitative(s). Cette dernière, appelée covariable, est souvent utilisée pour éliminer la variation externe de la variable dépendante.

Pour aller plus loin

Evrard Y., Pras B., et Roux E., *Market. Études et recherches en marketing*, Nathan, Paris, 2003.

Goupy J., *Introduction aux plans d'expérience*, Dunod, Paris, 2006.

Hair J. F., Anderson R. L., Black W. C., *Multivariate Data Analysis,* 4e éd., Prentice Hall International, Londres, 2006.

Malhotra *et al.*, *Études marketing avec SPSS*, Pearson Education, Paris, 2007.

Exercices

Exercice 1 Questions de recherche et type d'analyse de variance

Énoncé

Pour chacune des questions de recherche suivantes, trouvez le type d'analyse de variance approprié en spécifiant le nombre de facteurs avec leurs niveaux.

1. L'intention d'achat des consommateurs varie-t-elle en fonction de la couleur du packaging (rouge, vert ou bleu) ?
2. La CSP (5 catégories) a-t-elle un effet sur la qualité du service perçu ?
3. L'attitude vis-à-vis de la marque d'un produit de luxe varie-t-elle en fonction du pays d'origine de la marque (France, Espagne, Italie, États-Unis) et de son réseau de distribution (très sélectif ou non sélectif) ?
4. L'interaction entre le prix (élevé, moyen ou faible), la notoriété de la marque (forte ou faible) et la fréquence des contacts (forte ou faible) affecte-t-elle l'attitude vis-à-vis de la marque et l'intention d'achat de cette marque ?
5. Le niveau de prix (élevé ou faible), le conditionnement (familial, standard, mini) et l'attitude vis-à-vis des marques de lessive affectent-ils l'achat de lessive par les personnes âgées ?
6. Le style de la publicité (informative, humoristique, sexy) et l'attitude vis-à-vis de la marque ont-ils un impact sur l'intention d'achat d'un produit solaire de cette même marque ?
7. L'âge des clients (5 catégories) et la qualité du service perçue affectent-ils la satisfaction et la fidélisation des clients dans le secteur bancaire ?

Solution

1. ANOVA à un facteur, la couleur du packaging ayant trois niveaux (rouge, vert, bleu).

2. ANOVA à un facteur, la CSP ayant cinq niveaux.

3. ANOVA à deux facteurs, le pays d'origine de la marque et le réseau de distribution, lesquels ayant respectivement quatre niveaux (France, Espagne, Italie, États-Unis) et deux niveaux (très sélectif, non sélectif).

4. MANOVA à trois facteurs et deux variables expliquées. Les trois facteurs sont le prix, la notoriété de la marque et la fréquence des contacts, qui ont respectivement trois niveaux (élevé, moyen et faible), deux (forte ou faible) et deux (forte ou faible). Les deux variables expliquées sont l'attitude vis-à-vis de la marque et l'intention d'achat de la marque par les consommateurs.

5. ANCOVA à trois variables explicatives : deux variables qualitatives (prix et conditionnement) et une variable quantitative, covariable (attitude vis-à-vis des marques de lessive). Le prix et le conditionnement ont respectivement deux niveaux (élevé, faible) et trois (familial, standard, mini).

6. ANCOVA à deux variables explicatives : le style de la publicité (qui a trois niveaux, informative, humoristique, sexy) et la covariable attitude vis-à-vis de la marque.

7. MANCOVA à deux variables explicatives (l'âge [qui a 5 niveaux] et la qualité du service perçue [covariable]) et deux variables expliquées (la satisfaction et la fidélisation des clients).

EXERCICE 2 ÉTUDE DU POINT DE VENTE

Énoncé

Vous travaillez sur une enquête destinée à mieux comprendre les comportements d'achat des clients d'un magasin de chaussures. Vous cherchez à identifier ces clients et à connaître leur attitude à l'égard du point de vente. Vous avez collecté 400 réponses et vous voulez exploiter ces données (fichier « pointdevente » [1]).

Le gérant du magasin souhaite savoir :

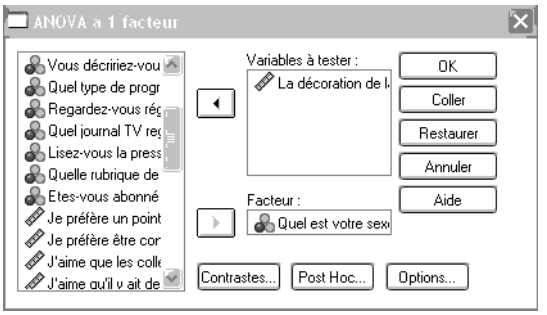
1. Si la décoration du magasin est plus importante pour les femmes que pour les hommes.
2. Si le montant dépensé par client est influencé par l'écoute régulière de médias (radio et TV).

Solution

1. Afin de savoir si la décoration du magasin a une influence en fonction du sexe des clients, il faut mener une ANOVA (voir figure 5.21). Allez dans le menu **Analyse** > **Comparer les moyennes** > **Anova à 1 facteur**. Indiquez la **décoration de la boutique** comme variable dépendante et le sexe comme variable indépendante puis cliquez sur OK.

Figure 5.21

Commande de l'ANOVA à 1 facteur.



Pour effectuer un test d'homogénéité (voir figure 5.22), cliquez sur **Options, Test d'homogénéité** puis sur **OK**.

Figure 5.22

Interprétation de l'ANOVA à 1 facteur.

Test d'homogénéité des variances

La décoration de la boutique est importante à mes yeux

Statistique de Levene	ddl1	ddl2	Signification
,367	1	398	,545

ANOVA

La décoration de la boutique est importante à mes yeux

	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Inter-groupes	,164	1	,164	,075	,784
Intra-groupes	866,273	398	2,177		
Total	866,438	399			

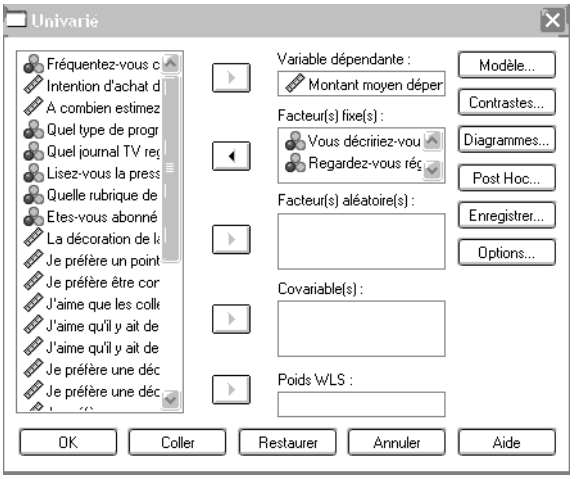
1. Vous trouverez ce fichier à l'adresse : http://www.pearsoneducation.fr.

Le test de Levene est significatif (0,78 > 0,05), l'hypothèse d'homogénéité des échantillons est donc acceptée. Les résultats de l'ANOVA attestent que le sexe n'a pas d'influence sur l'importance de la décoration (0,78).

2. Afin de savoir si le fait que les clients écoutent régulièrement des médias (radio et TV) a un impact sur le montant qu'ils dépensent, il faut faire une ANOVA. Allez dans le menu **Analyse > Modèle Linéaire Général > Univarié**. Choisissez comme variable dépendante **le montant moyen dépensé par mois**, et comme variable indépendante : **la fréquence d'écoute radio et TV** qui correspond aux questions : « Vous décririez-vous comme un auditeur régulier de radio ? » et « Regardez-vous régulièrement le journal télévisé ? »

Figure 5.23

Commande d'une ANOVA.



Faites ensuite un test d'homogénéité des variances (voir figure 5.24) : cliquez sur **Options, Test d'homogénéité** puis sur **OK**.

Figure 5.24

Interprétation d'une ANOVA.

Test d'égalité des variances des erreurs de Levene[a]

Variable dépendante : Montant moyen dépensé par mois

F	ddl1	ddl2	Signification
17,558	3	396	,000

Teste l'hypothèse nulle que la variance des erreurs de la variable dépendante est égale sur les différents groupes.

a. Plan : Ordonnée à l'origine+radio+journalTV+radio * journalTV

Ici le test de Levene (0,00) ne permet pas d'accepter l'hypothèse d'homogénéité des variances intragroupes. Les résultats de l'ANOVA ne sont donc pas valables et on ne peut pas dire que le fait de regarder régulièrement la TV ou la radio a un effet sur la consommation des clients dans le mois.

Exercice 3 Quel régime est le plus efficace ?

Énoncé

Un nutritionniste veut tester l'effet de différents régimes sur la perte de poids. Il souhaite évaluer l'effet de trois régimes sur des groupes d'individus. Le tableau suivant donne les résultats de ces régimes en nombre de kilogrammes perdus après un mois pour trois groupes d'individus ayant suivi les régimes.

Les groupes sont composés comme suit :

Groupe A : individus ayant suivi un régime hyperprotéiné ;

Groupe B : individus ayant suivi un régime d'association d'aliments ;

Groupe C : individus ayant suivi un régime hypocalorique.

Tableau 5.3 : Expérience sur les régimes

Groupe A	Groupe B	Groupe C
3	1	11
4	1	9
6	5	10
8	6	5
3	1	10
3	2	6
4	1	9
6	5	10
3		

Après avoir saisi ces données, faites une analyse de variance pour vérifier si les moyennes des trois groupes sont différentes.

Solution

Pour saisir les données, allez dans **Fichier** > **Nouveau** > **Données** (voir figure 5.25). Ensuite, dans **Affichage des variables**, rentrez le **Nom** des variables et leur **Étiquette**. Nous avons des données qualitatives (régime) et des données quantitatives (kilos perdus) : la colonne **Mesure** affiche **Nominales** et **Échelle.**

Il faut ensuite entrer les données dans la partie **Affichage des données** (voir figure 5.26).

Il faut ensuite commander une ANOVA à un facteur (voir figure 5.27). Allez dans le menu **Analyse** > **Comparer les moyennes** > **ANOVA à un facteur**. Choisissez comme variable dépendante **le nombre de kilos perdus** et comme variable indépendante **le type de régime**.

Demandez ensuite un test d'homogénéité des variances (voir figure 5.28). Cliquez sur **Options, Test d'homogénéité** puis sur **OK**.

Figure 5.25

Expérience sur les régimes : l'enregistrement des données.

Fichier Edition Affichage Données Transformer Analyse Graphes Outils Fenêtre Aide

	Nom	Type	Largeur	Décimales	Etiquette	Valeurs	Manquant	Colonnes	Aligner	Mesure
1	régime	Numérique	8	0		{1, hyperprotéin	Aucun	8	Droite	Nominale
2	kilos	Numérique	8	0		Aucun	Aucun	8	Droite	Echelle
3										

Figure 5.26

Expérience sur les régimes : l'enregistrement des données *(suite)*.

Fichier Edition Affichage Données

7 :

	régime	kilos
1	1	3
2	1	4
3	1	6
4	1	8
5	1	3
6	1	3
7	1	4
8	1	6
9	1	3
10	2	1
11	2	1
12	2	5
13	2	6
14	2	1
15	2	2
16	2	1

Figure 5.27

Expérience sur les régimes : commande de l'ANOVA.

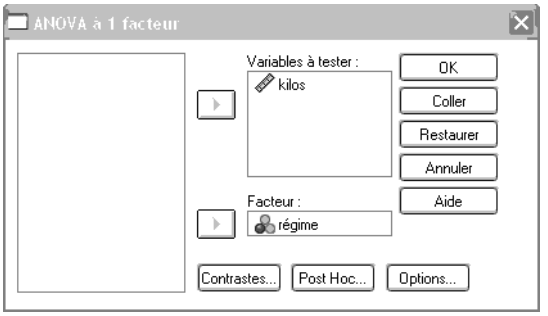


Figure 5.28

Expérience sur les régimes : interprétation des résultats de l'ANOVA.

Test d'homogénéité des variances

kilos

Statistique de Levene	ddl1	ddl2	Signification
,492	2	22	,618

ANOVA

kilos

	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Intergroupes	153,818	2	76,909	18,548	,000
Intragroupes	91,222	22	4,146		
Total	245,040	24			

Le test de Levene est significatif (0,52), l'hypothèse d'homogénéité des échantillons est donc acceptée. Nous pouvons procéder à l'analyse des résultats de l'ANOVA.

Le type de régime a un effet significatif sur la perte de poids. Cependant, à ce stade, nous ne connaissons pas le type de régime le plus efficace. Il faut demander un test de différence de moyenne (test post hoc) [voir figure 5.29].

Allez dans le menu **Analyse > Comparer les moyennes > ANOVA à un facteur**, cliquez sur le bouton **Post Hoc** puis choisissez un test (ici, le test de Duncan).

Figure 5.29

Expérience sur les régimes : interprétation des résultats de l'ANOVA (suite).

Tests post hoc

Sous-ensembles homogènes

kilos

Duncan[a,b]

régime	N	Sous-ensemble pour alpha = .05	
		1	2
associations	8	2,75	
hyperprotéiné	9	4,44	
hypocalorique	8		8,75
Signification		,104	1,000

Les moyennes des groupes des sous-ensembles homogènes sont affichées.

a. Utilise la taille d'échantillon de la moyenne harmonique = 8,308.

b. Les effectifs des groupes ne sont pas égaux. La moyenne harmonique des effectifs des groupes est utilisée. Les niveaux des erreurs de type I ne sont pas garantis.

Les résultats de l'ANOVA montrent que le régime hypocalorique est le plus efficace. En effet, la moyenne des kilos perdus par les personnes qui ont suivi ce régime est significativement plus importante que les autres.

Chapitre 6

La régression linéaire

Il est fréquent d'observer des phénomènes où l'on peut penser qu'il existe une liaison entre deux variables. Par exemple, l'âge d'une voiture et son kilométrage varient généralement dans le même sens. Ce lien n'est cependant pas absolu : comment mesurer l'intensité de la relation entre ces deux variables ? Le modèle de régression linéaire a pour objectif d'expliquer la variation d'un phénomène mesurable (variable dépendante quantitative) par celle d'un ou de plusieurs autres (variables quantitatives). La régression linéaire simple ou multiple estime les coefficients de l'équation linaire impliquant cette ou ces variables indépendantes, qui évaluent le mieux la valeur de la variable dépendante.

Cette méthode est largement utilisée en marketing, par exemple pour expliquer les variations des ventes, de préférence de marques, produits ou services.

Avant de présenter l'analyse de régression, nous abordons le coefficient de corrélation qui constitue la base conceptuelle de la régression.

1 La corrélation linéaire

La corrélation linéaire est une statistique largement utilisée car elle synthétise l'importance de la relation entre deux variables métriques. Le tableau 6.1 montre bien que le coefficient de corrélation est le test statistique pour mesurer le lien entre deux variables quantitatives.

Tableau 6.1 : Rappel sur la nature des variables et le type d'analyse

Nature des variables	Type d'analyse	Test statistique
Qualitatives	Tri croisé	Khi-deux
Qualitatives et quantitatives	ANOVA	Test *F*
Quantitatives	Régression	Coefficient de corrélation

Après avoir présenté les principes de la corrélation, nous procédons à une démonstration avec la réalisation d'une corrélation multiple avec le logiciel SPSS.

1.1 Les principes de la corrélation linéaire

Le **coefficient de corrélation de Pearson** est une mesure d'association qui permet d'établir si deux variables mesurées sur le même ensemble d'observations varient de façon analogue ou non.

La corrélation « r » est égale à la covariance divisée par le produit des écarts types de x et y :

$$r = covXY/Sx\ Sy$$

Cette corrélation correspond également au coefficient de régression (b) divisé par l'écart type de la variable dépendante :

$$r = b/Sy$$

Une corrélation proche de 1 ou de –1 en valeur absolue signifie que deux variables sont liées entre elles et peuvent s'expliquer mutuellement. Lorsque r est proche de 0, il y a une faible corrélation. Si r est proche de +1, cela veut dire que les deux variables varient dans le même sens. Si r est proche de –1, cela signifie que les deux variables varient en sens inverse l'une de l'autre.

1.2 La réalisation d'une corrélation linéaire

Avant de réaliser une corrélation linéaire, il faut s'assurer que les variables à tester sont bien quantitatives. En effet, comme nous l'avons vu au tableau 6.1, le coefficient de corrélation ne fonctionne que pour des variables métriques.

SPSS

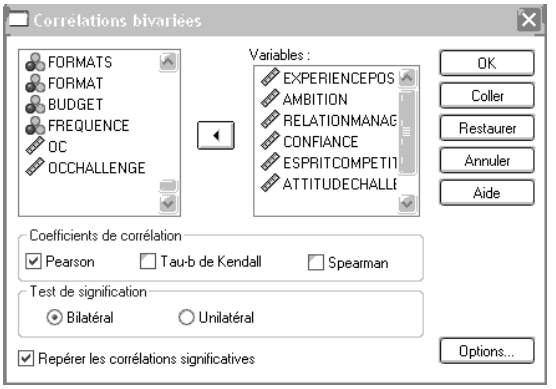
Étude des liens entre diverses caractéristiques des vendeurs et leur attitude à l'égard des challenges

Nous cherchons à savoir s'il existe une relation entre des caractéristiques du vendeur telles que l'âge, l'ancienneté dans l'entreprise et dans le poste, l'ambition et l'attitude à l'égard des challenges de vente. Nous réalisons donc une corrélation linéaire sur toutes ces variables quantitatives.

Ouvrez le fichier « challenge » [1]. Allez dans le menu **Analyse > Corrélation > Bivariée.** La boîte de dialogue de la figure 6.1 apparaît.

Figure 6.1

Commande d'une corrélation.



Gardez le coefficient de Pearson coché. Transférez les variables en les sélectionnant chacune à leur tour et en cliquant sur les flèches. Faites **OK**. Les résultats apparaissent (voir figure 6.2).

Figure 6.2

Interprétation d'une corrélation.

Corrélations

		EXPERIENCEPOSTE	AMBITION	RELATIONMANAGER	CONFIANCE	ESPRITCOMPETITION	ATTITUDECHALLENGE
EXPERIENCEPOSTE	Corrélation de Pearson	1	-,229**	-,010	,109**	-,170**	-,204**
	Sig. (bilatérale)		,000	,787	,004	,000	,000
	N	733	697	687	700	695	690
AMBITION	Corrélation de Pearson	-,229**	1	,314**	,144**	,601**	,476**
	Sig. (bilatérale)	,000		,000	,000	,000	,000
	N	697	705	685	696	691	672
RELATIONMANAGER	Corrélation de Pearson	-,010	,314**	1	,301**	,293**	,318**
	Sig. (bilatérale)	,787	,000		,000	,000	,000
	N	687	685	695	685	682	663
CONFIANCE	Corrélation de Pearson	,109**	,144**	,301**	1	,135**	,122**
	Sig. (bilatérale)	,004	,000	,000		,000	,001
	N	700	696	685	708	693	677
ESPRITCOMPETITION	Corrélation de Pearson	-,170**	,601**	,293**	,135**	1	,713**
	Sig. (bilatérale)	,000	,000	,000	,000		,000
	N	695	691	682	693	704	672
ATTITUDECHALLENGE	Corrélation de Pearson	-,204**	,476**	,318**	,122**	,713**	1
	Sig. (bilatérale)	,000	,000	,000	,001	,000	
	N	690	672	663	677	672	701

**. La corrélation est significative au niveau 0.01 (bilatéral).

Les résultats indiquent le coefficient de corrélation et la signification (Sig.). Si Sig. < 0,05, on peut dire qu'il existe une corrélation entre les deux variables au seuil de 0,05. Le signe ** indique que la corrélation est significative au seuil de 0,01.

Dans cet exemple, nous observons que l'esprit de compétition et l'attitude à l'égard des challenges de vente sont fortement liés (0,71 ; $p < 0,01$).

À ce stade, nous ne pouvons cependant pas dire si c'est l'esprit de compétition qui a un impact sur l'attitude à l'égard du challenge ou bien l'inverse. C'est grâce à la régression linéaire que nous pouvons expliquer le sens de la relation entre ces deux variables.

1. Vous trouverez ce fichier à l'adresse : http://www.pearsoneducation.fr.

2 La régression linéaire

La régression linéaire vise à expliquer une variable dépendante par une ou un ensemble de variables indépendantes quantitatives. Lorsque le problème implique une seule variable indépendante, la technique statistique est appelée régression simple. Lorsque le problème implique plusieurs variables indépendantes, il s'agit d'une régression multiple. La régression est utilisée pour l'explication et la prédiction.

Les principes et conditions d'application de la régression simple sont exposés avant d'aborder ceux de la régression multiple. Ces deux techniques sont chacune illustrées par des exemples d'applications.

2.1 La régression linéaire simple

La régression vise à estimer ou prédire la valeur d'une variable à partir d'une seule autre. Par exemple, on peut expliquer la consommation de SMS par l'âge du consommateur.

Dans une régression simple, les valeurs de la variable dépendante (Y) sont estimées à partir de la variable indépendante (X) par équation linéaire :

$$Yi = aC + bXi + e$$

Où *Yi* est la valeur estimée de Y, *b* est la pente (coefficient de régression) et *a* la constante.

Les conditions d'application de la régression

Le modèle de la régression pose un certain nombre d'hypothèses lors de l'estimation des paramètres et des tests d'hypothèses. Ces conditions d'application de la régression sont :

- la linéarité du phénomène mesuré ;
- la variance constante du terme d'erreur ou homoscédasticité ;
- l'indépendance des termes d'erreur ;
- la normalité de la distribution du terme d'erreur.

La linéarité est importante car le concept de corrélation est fondé sur une relation linéaire. La linéarité d'une relation bivariée est vérifiée par l'examen des résidus.

L'homoscédasticité est vérifiée par l'examen des résidus ou par un simple test statistique. Le logiciel SPSS fournit le test d'homogénéité de Levene, qui mesure l'égalité des variances pour une seule paire de variables. Son utilisation est souvent recommandée.

L'indépendance des termes d'erreur est une autre condition de l'analyse de régression multiple. Outre l'examen du graphique des résidus, cette hypothèse peut aussi être validée par le test de Durbin-Watson.

La normalité de la distribution du terme d'erreur (voir chapitre 2).

L'interprétation des résultats de la régression

Les résultats de la régression se lisent grâce aux indices suivants :

- **R** : le **coefficient de corrélation multiple** est un indice standardisé variant de –1 à +1, indiquant la force de la relation entre l'ensemble des variables indépendantes et la

variable dépendante. Plus la corrélation est élevée, plus la relation linéaire entre les variables indépendantes et la variable dépendante est élevée.

- **R^2** : la corrélation multiple au carré, appelée **coefficient de détermination**, est un indice de la part de variance de la variable dépendante expliquée par les variables indépendantes qui sont dans l'équation. Il donne ainsi la part de variance de la variable expliquée par la variable indépendante.
- **Le Bêta** : ce **coefficient standardisé** permet de comparer la contribution de chaque variable puisqu'il s'agit du coefficient de régression ramené sur une échelle standard (entre –1 et +1).
- **Le test *F***: sa valeur indique si la variance ou l'ajout de variance expliquée est significative, c'est-à-dire si, quelle que soit la force de la relation entre les variables indépendantes et la variable dépendante, cette relation est susceptible d'exister dans la population et n'est pas due simplement au hasard de l'échantillonnage.
- **Le test *t***: sa valeur doit être plus grande que 2 (1,96) pour être significative (notée** à $p < 0{,}05$). Elle indique si chacun des coefficients des variables présentes dans l'équation est significatif.

Focus 6.1 Concomitance et corrélation

Concomitance et corrélation ne veulent pas dire obligatoirement relation de cause à effet. Il faut d'abord comprendre le lien de causalité entre la variable explicative et la ou les variables expliquées, vérifier expérimentalement la solidité du lien (*via* les méthodes de régression), et seulement alors s'en servir en explicatif ou en prévisionnel. Le risque sinon est de mettre en évidence une relation forte entre deux grandeurs n'ayant aucune relation de cause à effet, mais simplement reliées toutes les deux pour des raisons très différentes à une même troisième.

Par exemple, on cite fréquemment l'exemple de la bonne corrélation entre le nombre de meurtres par an en Grande-Bretagne et la consommation de chocolat ; de là à en conclure que la consommation de chocolat rend agressif ! (alors que les deux varient en fonction de la population, et si on neutralise cet effet, il n'y a aucune corrélation, à population fixée, entre la criminalité et la consommation de chocolat !). Ce risque est particulièrement présent lorsque l'on adopte des méthodes de type pas à pas, en introduisant les variables « explicatives » uniquement en fonction de critères de performance et non pas en analysant sur un plan conceptuel la relation de cause à effet.

SPSS

Étude de la relation entre l'esprit de compétition et l'attitude envers le challenge

Nous cherchons à savoir si l'esprit de compétition du vendeur influence son attitude à l'égard du challenge de vente.

Pour obtenir une régression linéaire simple, allez dans le menu **Analyse > Régression > Linéaire**. La boîte de dialogue de la figure 6.3 apparaît. Transférez les variables en les sélectionnant chacune à leur tour et en cliquant sur les flèches. Mettez la variable à expliquer dans **Variable dépendante**, la variable explicative dans **Variables explicatives.**

Le premier tableau récapitule les variables explicatives prises en compte dans le modèle. Ici, il n'y a qu'une seule variable puisque nous travaillons sur une régression simple.

Le troisième tableau indique si le modèle est significatif ou non. Dans ce cas-ci, le modèle obtenu est significatif ($p < 0{,}05$), le lien entre l'esprit de compétition et l'attitude du vendeur à l'égard des challenges de vente est significatif ($t = 26{,}34 > 2$) et positif (R^2 ou coefficient standardisé de 0,50) [voir le premier tableau de la figure 6.4].

Figure 6.3

Commande d'une régression simple.

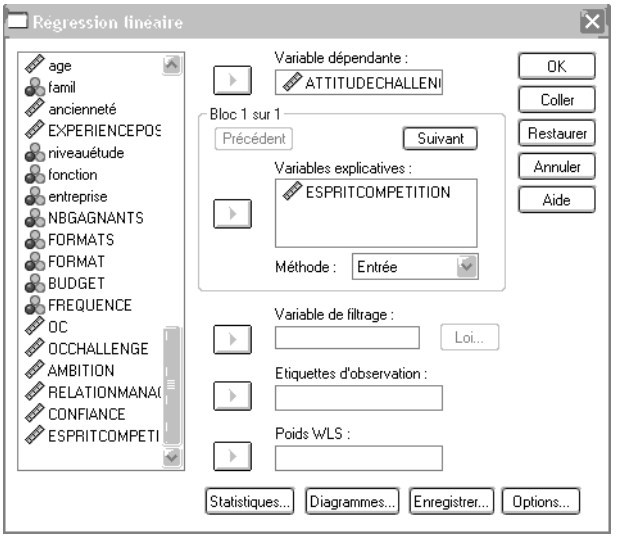


Figure 6.4

Interprétation d'une régression simple.

Variables introduites/éliminées[b]

Modèle	Variables introduites	Variables éliminées	Méthode
1	ESPRITCOMPETITION[a]	.	Introduire

a. Toutes variables requises introduites

b. Variable dépendante : ATTITUDECHALLENGE

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,713[a]	,509	,508	,74689

a. Valeurs prédites : (constantes), ESPRITCOMPETITION

Figure 6.5

Interprétation d'une régression simple (suite).

ANOVA[b]

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	387,196	1	387,196	694,089	,000[a]
	Résidu	373,758	670	,558		
	Total	760,955	671			

a. Valeurs prédites : (constantes), ESPRITCOMPETITION

b. Variable dépendante : ATTITUDECHALLENGE

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification
		B	Erreur standard	Bêta		
1	(constante)	,394	,093		4,246	,000
	ESPRITCOMPETITION	,828	,031	,713	26,346	,000

a. Variable dépendante : ATTITUDECHALLENGE

Focus 6.2 De la nécessité d'observer les données au préalable

Tout travail de type recherche de corrélation et de modélisation commence par une séance approfondie de statistique descriptive. Avant de faire des calculs de régression, regardons attentivement les données. Il faut en particulier se méfier des points aberrants, susceptibles de *« tirer »* les coefficients de régression, ou d'un nuage de points organisé en deux paquets orientés suivant deux directions, ou d'autres cas de ce type.

Ce travail se fait par l'examen des résidus comme nous allons maintenant le découvrir.

2.2 L'EXAMEN DES RÉSIDUS

L'estimation réalisée par l'équation de régression n'atteint habituellement pas l'exactitude complète. D'un point de vue géométrique, les points des données ne se retrouvent pas sur la ligne droite spécifiée par l'équation de régression. Les résidus représentent les différences sur les variables prédites ; ils constituent un indicateur de performance de la droite de régression.

L'examen de ces résidus sert à estimer l'exactitude des estimations. Il est donc recommandé de demander une analyse des résidus avec des graphiques qui fournissent des aperçus utiles pour s'assurer que les hypothèses fondamentales et la qualité du modèle de régression ont bien été respectées.

L'hypothèse de **distribution normale du terme d'erreur** est vérifiée par l'observation du graphique des résidus. L'hypothèse d'une **valeur constante de la variance du terme d'erreur** (**homoscédasticité**) est validée à l'aide du graphique des résidus en fonction des valeurs estimées de la variable dépendante. Si la configuration n'est pas aléatoire, la variance du terme d'erreur n'est pas constante. La variation des variances des termes d'erreur doit être comprise entre –3 et +3. Ce graphique sert aussi à vérifier une autre condition importante : l'**absence de corrélation ou indépendance entre les termes d'erreur.**

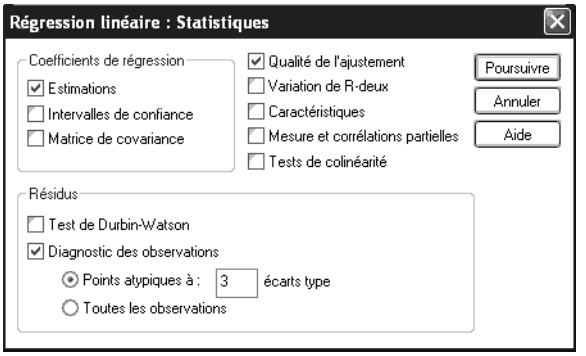
L'exemple suivant montre comment demander un examen des résidus avec le logiciel SPSS et comment interpréter les résultats de ce diagnostic.

SPSS

Pour obtenir l'examen des résidus, retournez à la boîte de dialogue (**Analyse > Régression > Linéaire**), cliquez sur **Statistiques** et, dans l'encadré **Résidus**, cochez **Diagnostic des observations** (voir figure 6.6).

Figure 6.6

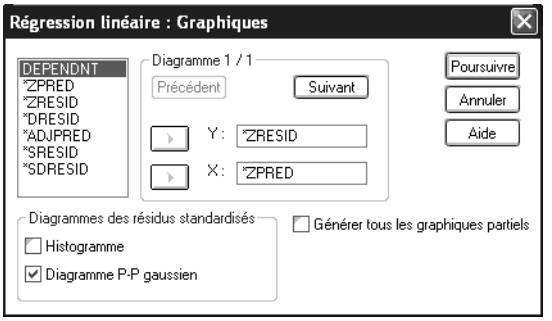
Demande d'un tableau des résidus.



Cliquez ensuite sur **Poursuivre** et, pour avoir un graphique des résidus, cliquez sur le bouton **Diagrammes**. La boîte de dialogue de la figure 6.7 apparaît.

Figure 6.7

Demande d'un diagramme des résidus *(suite)*.



Pour commander un diagramme de résidus standardisés (*ZRESID) contre les valeurs prédites standardisées (ZPRED), il faut transférer avec les flèches *ZRESID dans la case face à Y et *ZPRED dans la case face à X. Cochez **Diagramme P-P gaussien** afin d'avoir la droite de régression. Cliquez ensuite sur **Poursuivre** pour revenir à la boîte de dialogue précédente et enfin sur **OK**.

Figure 6.8

Diagnostic des observations.

Diagnostic des observations[a]

Numéro de l'observation	Résidu standardisé	ATTITUDEC HALLENGE	Prévision	Résidu
61	4,166	4,33	1,2220	3,11135
202	3,057	4,33	2,0498	2,28351
337	3,612	4,33	1,6359	2,69743

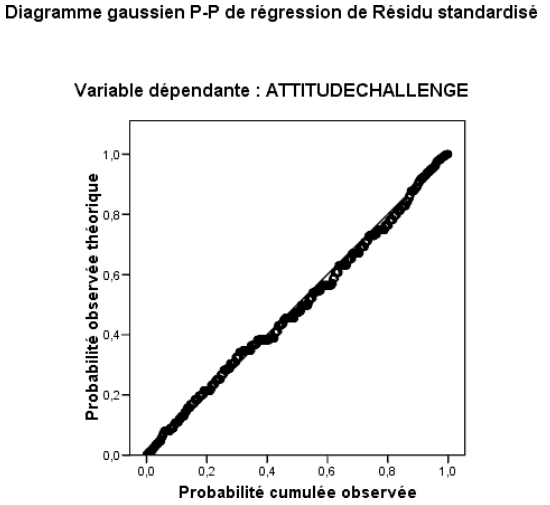
a. Variable dépendante : ATTITUDECHALLENGE

Le tableau Diagnostic des observations montre que les observations 61, 202 et 337 sortent de l'intervalle [–3 ; +3] avec un score moyen de 4,33 pour l'attitude à l'égard du challenge. Le résidu standardisé est supérieur à 4 pour l'observation 61 et 3 pour les deux dernières.

Afin d'éliminer ces observations éloignées de la droite de régression, il faut aller dans le menu **Données > Sélectionnez des observations**. Cliquez sur le bouton **Selon une condition logique** (voir figure 6.9).

Figure 6.9

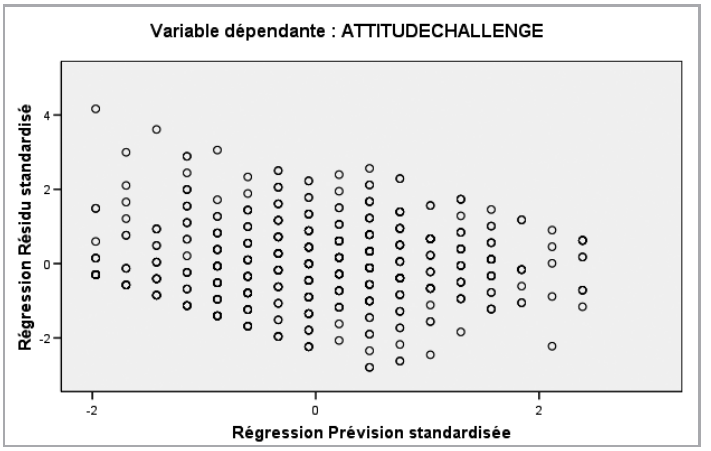
Diagnostic des résidus.



Le graphique de répartition des résidus par rapport à une répartition normale montre que la majorité des résidus est alignée.

Figure 6.10

Nuage de points des résidus standardisés.



Le nuage de points édité des résidus standardisés en rapport avec les valeurs prédites standardisées ne fait apparaître aucun modèle particulier, ce qui confirme l'**hypothèse de valeur constante de la variance du terme d'erreur (homoscédasticité) et d'indépendance des termes d'erreur.**

D'autres diagrammes, comme l'histogramme des résidus standardisés, sont aussi à examiner. Idéalement, la distribution doit être normale.

2.3 La régression linéaire multiple

La régression multiple est une extension de la régression simple où la variable dépendante est régressée sur un ensemble de variables. Elle sert à analyser la relation entre une variable dépendante qualitative et plusieurs variables indépendantes quantitatives. Chaque variable indépendante est évaluée par la procédure de régression de façon à maximiser la prédiction de la variable expliquée.

Cette technique multivariée est la plus utilisée pour prédire et expliquer. Dans le cas de la prédiction, l'objectif est de maximiser le pouvoir prédictif des variables indépendantes. Il est aussi possible de comparer des variables indépendantes dans leur pouvoir explicatif. Dans le cas de l'explication, la régression sert à déterminer l'importance relative de chaque variable indépendante par sa magnitude et sa direction. Par exemple, le nombre de SMS peut dépendre de l'âge du consommateur, de son revenu et de ses consommations téléphoniques.

La régression cherche la combinaison de poids (b) pour les variables indépendantes (Xi) qui amènerait les valeurs de Y prédites par l'équation aussi près que possible des valeurs de Y mesurées :

$$Yi = a + b1X1 + b2X2 + \ldots + bnXn$$

C'est un indice de la relation entre les valeurs prédites et les valeurs mesurées.

Les méthodes de sélection des variables de régression

La sélection d'une méthode permet de spécifier la manière dont les variables indépendantes sont entrées dans l'analyse.

Voici ces différentes méthodes :

- **entrée (par défaut)** : toutes les variables d'un bloc sont introduites en une seule opération ;
- **pas à pas** : les variables indépendantes sont ajoutées à l'équation une par une et peuvent être enlevées subséquemment si elles ne contribuent plus significativement à la régression. Le processus s'arrête lorsqu'aucune variable ne peut plus être introduite ou éliminée ;
- **éliminer bloc** : toutes les variables dans un bloc sont supprimées en une seule étape ;
- **descendante** : toutes les variables sont entrées initialement dans l'équation et sont ensuite éliminées une à une. La variable ayant la plus petite corrélation avec la variable dépendante est d'abord étudiée pour l'élimination. Si elle est éliminée par le modèle, la prochaine variable avec le plus petit coefficient de corrélation est étudiée, jusqu'à ce qu'aucune variable ne satisfasse plus au critère d'élimination ;
- **ascendante** : les variables sont introduites séquentiellement une par une. Si la première variable est introduite dans l'équation, la variable explicative ne figurant pas dans l'équation et présentant la plus forte corrélation partielle est considérée ensuite. La procédure s'arrête lorsqu'il ne reste plus de variables satisfaisant le critère d'introduction.

La méthode pas à pas est une **combinaison des méthodes descendantes et ascendantes**, elle est généralement recommandée comme étant la meilleure méthode.

Toutes les variables doivent respecter le critère de tolérance pour être entrées dans l'équation, quelle que soit la méthode d'entrée spécifiée. Le niveau de tolérance par défaut est 0,0001. Une variable n'est pas entrée si elle fait passer la tolérance d'une autre variable déjà entrée dans le modèle en dessous du seuil de tolérance.

Les conditions de la régression multiple

La régression multiple est complexifiée par la présence de **multicolinéarité.** En effet, la majorité des études mettent en jeu des variables explicatives qui sont corrélées. Une méthode simple pour détecter une trop grande corrélation entre variables indépendantes consiste à demander des tests de colinéarité : **tolérance** et **facteur d'inflation de la variance** (VIF).

La tolérance est définie comme la part de variabilité de la variable indépendante qui n'est pas expliquée par une ou d'autres variables indépendantes. Une tolérance élevée correspond à un faible degré de colinéarité. Le seuil de 0,3 est recommandé. À l'inverse, le seuil du facteur d'inflation de la variance (VIF) doit être faible : < 3.

SPSS

Étude de la relation entre l'esprit de compétition, l'ambition, la relation avec le manager et l'attitude des vendeurs envers les challenges de vente

Nous cherchons à savoir si l'esprit de compétition, l'ambition et la relation du vendeur avec son manager influencent son attitude à l'égard des challenges de vente. Nous souhaitons déterminer, parmi ces variables explicatives, celle qui explique le mieux l'attitude à l'égard du challenge de vente.

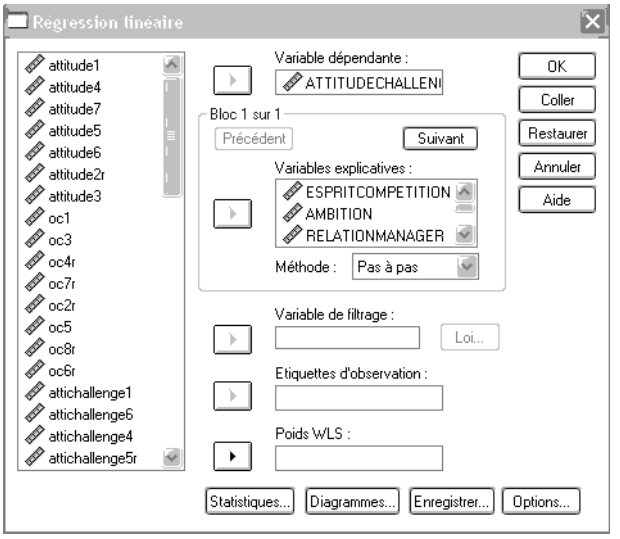
Pour obtenir une régression linéaire multiple, allez dans le menu **Analyse > Régression > Linéaire**. La boîte de dialogue de la figure 6.11 apparaît.

Transférez les variables en les sélectionnant chacune à leur tour et en cliquant sur les flèches. La variable à expliquer dans **Variable dépendante**, les variables explicatives dans **Variables explicatives**.

La méthode de sélection des variables **pas à pas** est choisie eu égard à notre choix de recherche.

Figure 6.11

Commande d'une régression multiple.



Cliquez ensuite sur **Statistiques** et demandez les **Tests de colinéarité**.

Cliquez ensuite sur **Poursuivre** et **OK** (voir figure 6.12).

Figure 6.12

Commande de tests de colinéarité.

Les résultats de la régression multiple apparaissent dans l'onglet Résultats (voir figures 6.13 et 6.14).

Ce premier tableau présente les variables introduites : l'esprit de compétition et la relation avec le manager. Nous remarquons ici que l'ambition n'est pas prise en compte car cette variable ne contribue pas significativement à la régression.

Les deux variables prises en compte expliquent 51 % de l'attitude à l'égard du challenge (R^2 ajusté). Le tableau ANOVA atteste que les deux modèles sont significatifs (Signification = 0,00).

Nous lisons ensuite les résultats du test de colinéarité (voir figure 6.15).

Figure 6.13

Interprétation des résultats d'une régression multiple.

Variables introduites/éliminées[a]

Modèle	Variables introduites	Variables éliminées	Méthode
1	ESPRITCOMPETITION	.	Pas à pas (critère : Probabilité de F pour introduire <= ,050, Probabilité de F pour éliminer >= ,100).
2	RELATIONMANAGER	.	Pas à pas (critère : Probabilité de F pour introduire <= ,050, Probabilité de F pour éliminer >= ,100).

a. Variable dépendante : ATTITUDECHALLENGE

Figure 6.14

Interprétation des résultats d'une régression multiple (suite).

Récapitulatif du modèle[c]

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,707[a]	,499	,498	,75202
2	,717[b]	,514	,512	,74167

a. Valeurs prédites : (constantes), ESPRITCOMPETITION

b. Valeurs prédites : (constantes), ESPRITCOMPETITION, RELATIONMANAGER

c. Variable dépendante : ATTITUDECHALLENGE

ANOVA[c]

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	361,402	1	361,402	639,053	,000[a]
	Résidu	362,503	641	,566		
	Total	723,905	642			
2	Régression	371,854	2	185,927	338,001	,000[b]
	Résidu	352,050	640	,550		
	Total	723,905	642			

a. Valeurs prédites : (constantes), ESPRITCOMPETITION

b. Valeurs prédites : (constantes), ESPRITCOMPETITION, RELATIONMANAGER

c. Variable dépendante : ATTITUDECHALLENGE

Figure 6.15

Interprétation des résultats d'une régression multiple : tests de colinéarité.

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Statistiques de colinéarité	
		B	Erreur standard	Bêta	t	Signification	Tolérance	VIF
1	(constante)	,412	,096		4,280	,000		
	ESPRITCOMPETITION	,824	,033	,707	25,280	,000	1,000	1,000
2	(constante)	,000	,134		-,003	,998		
	ESPRITCOMPETITION	,783	,034	,671	23,360	,000	,920	1,087
	RELATIONMANAGER	,157	,036	,125	4,359	,000	,920	1,087

a. Variable dépendante : ATTITUDECHALLENGE

Tolérances et facteurs d'inflation de la variance (VIF) sont proches de 1, largement dans les limites recommandées (tolérance > 0,3 et VIF < 3,3). Les variables explicatives sont donc peu corrélées entre elles, ce qui est un indice de qualité du modèle.

Comme nous pouvons le voir (voir figure 6.16), la variable ambition a été exclue car elle a de mauvaises statistiques de colinéarité.

Figure 6.16

Interprétation des résultats d'une régression multiple : variables exclues.

Variables exclues[c]

Modèle		Bêta dans	t	Signification	Corrélation partielle	Statistiques de colinéarité		
						Tolérance	VIF	Tolérance minimale
1	AMBITION	,083[a]	2,345	,019	,092	,626	1,598	,626
	RELATIONMANAGER	,125[a]	4,359	,000	,170	,920	1,087	,920
2	AMBITION	,058[b]	1,626	,104	,064	,606	1,649	,606

a. Valeurs prédites dans le modèle : (constantes), ESPRITCOMPETITION

b. Valeurs prédites dans le modèle : (constantes), ESPRITCOMPETITION, RELATIONMANAGER

c. Variable dépendante : ATTITUDECHALLENGE

Diagnostics de colinéarité[a]

Modèle	Dimension	Valeur propre	Index de conditionnement	Proportions de la variance		
				(constante)	ESPRITCOMPETITION	RELATIONMANAGER
1	1	1,951	1,000	,02	,02	
	2	,049	6,330	,98	,98	
2	1	2,910	1,000	,01	,01	,01
	2	,060	6,937	,07	,95	,23
	3	,030	9,911	,92	,04	,76

a. Variable dépendante : ATTITUDECHALLENGE

Nous pouvons conclure que le modèle 2 est satisfaisant, car il explique 51 % de l'attitude à l'égard du challenge (R^2 ajusté). Il est significatif (voir tableau ANOVA, figure 6.14), les coefficients de la pente de régression sont significatifs et il n'y a pas de problème de colinéarité (voir tableau coefficients, figure 6.15).

L'esprit de compétition et la relation du vendeur avec son manager sont deux déterminants de l'attitude du vendeur à l'égard des challenges de vente.

Résumé

La corrélation sert à mesurer la force de l'association de deux variables quantitatives. Le coefficient de corrélation linéaire mesure la relation linéaire entre les deux variables quantitatives.

La régression utilise la présence de cette relation pour prédire les valeurs de la variable dépendante à partir d'une variable indépendante. L'objectif est donc d'estimer ou de prédire une variable à partir d'une autre grâce à une équation de régression.

La régression simple sert à tester l'effet d'une seule variable indépendante sur une variable dépendante. La force de la relation est mesurée par le coefficient de détermination R^2. La régression multiple implique au moins deux variables indépendantes et une variable dépendante. La signification de l'équation de régression globale est testée grâce au test *t*. Les graphiques des résidus servent à vérifier la pertinence des hypothèses sous-jacentes et l'ajustement du modèle de régression

Pour aller plus loin

Malhotra N., Decaudin J. M., Bouguerra A., *Études marketing avec SPSS*, 5e éd., Pearson Education, Paris, 2007.

Hair J. F., Anderson R. E., Tatham R. L., Black W. C., *Multivariate Data Analysis*, Prentice Hall International, New Jersey, 2007.

Evrard Y., Pras B., Roux E., *Market. Études et recherche en marketing*, Nathan, Paris, 2003.

Exercices

Exercice 1 Étude du point de vente

Énoncé

Une étude sur la clientèle d'un magasin vise à mieux comprendre les clients et notamment les variables liées à l'intention d'achat. Voici les résultats de la corrélation multiple entre le montant moyen dépensé par mois, l'intention d'achat, le niveau d'études, la taille du foyer et le niveau de revenus du foyer (voir figure 6.17).

1. Quelles variables sont le plus reliées au montant moyen dépensé par mois ?
2. Ces variables sont-elles différentes de celles qui sont reliées à l'intention d'achat dans le point de vente ?

Figure 6.17

Interprétation des résultats d'une corrélation multiple.

Corrélations

		Montant moyen dépensé par mois	Intention d'achat dans le point de vente	Quel est votre niveau d'étude?	Taille du foyer	Revenus du foyer
Montant moyen dépensé par mois	Corrélation de Pearson	1	,333**	,666**	,038	,833**
	Sig. (bilatérale)		,000	,000	,445	,000
	N	400	400	400	400	400
Intention d'achat dans le point de vente	Corrélation de Pearson	,333**	1	,502**	,018	,283**
	Sig. (bilatérale)	,000		,000	,719	,000
	N	400	400	400	400	400
Quel est votre niveau d'étude?	Corrélation de Pearson	,666**	,502**	1	-,064	,635**
	Sig. (bilatérale)	,000	,000		,205	,000
	N	400	400	400	400	400
Taille du foyer	Corrélation de Pearson	,038	,018	-,064	1	-,002
	Sig. (bilatérale)	,445	,719	,205		,974
	N	400	400	400	400	400
Revenus du foyer	Corrélation de Pearson	,833**	,283**	,635**	-,002	1
	Sig. (bilatérale)	,000	,000	,000	,974	
	N	400	400	400	400	400

**La corrélation est significative au niveau 0.01 (bilatéral).

Solution

1. Les variables les plus reliées au montant moyen dépensé par mois sont le niveau d'études (0,66) et le revenu du foyer (0,83). L'intention d'achat est plus faiblement corrélée au montant moyen dépensé par mois. Le signe ** indique que la corrélation est significative au seuil de 0,01.

La taille du foyer n'est pas reliée au montant moyen dépensé par mois. En effet, la corrélation de 0,03 n'est pas significative, il n'y a pas de signe ** à l'intersection de « Taille du foyer » et de « Montant moyen dépensé par mois ».

2. Comme pour le montant moyen dépensé par mois, l'intention d'achat est aussi liée au niveau d'études et n'est pas liée à la taille du foyer. Toutefois, à la différence du montant moyen dépensé par mois, l'intention d'achat et le niveau de revenus du foyer ne sont pas fortement corrélés (0,28).

Exercice 2 Les déterminants de la moyenne des étudiants

Énoncé

On a demandé à des étudiants d'une classe d'évaluer la qualité de l'enseignement en utilisant une échelle de notation en 5 points (1 = médiocre, 5 = excellent). Nous avons aussi relevé la moyenne et le jour d'absence du trimestre des étudiants de la classe.

1. Enregistrez les données récoltées dans le tableau sous SPSS au tableau 6.2.
2. Ces variables sont-elles corrélées ?
3. Effectuez une analyse par régression multiple de la qualité perçue de l'enseignement et de l'absentéisme durant le trimestre sur la moyenne du trimestre. Interprétez les coefficients de régression. La régression est-elle significative ? Que concluez-vous ?

Tableau 6.2 : Données récoltées

Étudiant	Moyenne du trimestre	Qualité perçue de l'enseignement	Absentéisme en jours par trimestre
1	5	4	10
2	7	4	2
3	15	3	0
4	11	2	0
5	16	3	0
6	12	2	1
7	11	3	0
8	14	4	0
9	10	4	0
10	14	4	0
11	11	3	0
12	9	4	1
13	9	2	2
14	11	4	0
15	10	4	0
16	7	2	1
17	14	4	0
18	15	4	0
19	11	3	0
20	14	4	0

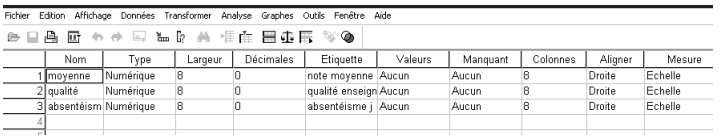
Étudiant	Moyenne du trimestre	Qualité perçue de l'enseignement	Absentéisme en jours par trimestre
21	12	3	0
22	11	4	0
23	9	3	0
24	8	2	2
25	11	4	0
26	10	3	0
27	14	4	0
28	12	4	0

Solution

1. Pour rentrer ces données sous SPSS, allez dans **Fichier** > **Nouveau** > **Données**. Ensuite, dans **Affichage des variables** entrez le **Nom** des variables et leur **Etiquette**. Nous avons ici des données quantitatives et donc la **Mesure** sélectionnée est **Echelle** (voir figure 6.18).

Figure 6.18

Rappel de la procédure pour rentrer les données sous SPSS.



Il faut ensuite entrer les données dans la partie **Affichage des données**. Chaque ligne correspond à la réponse d'un étudiant. Nous avons ainsi un tableau de trois colonnes et 28 lignes pour les 28 étudiants de la classe (voir figure 6.19).

2. Afin de savoir si les variables sont corrélées, il faut demander une corrélation entre les trois variables : qualité de l'enseignement perçue, absentéisme des élèves et moyenne du trimestre.

Allez dans **Analyse** > **Corrélation** > **Bivariée**. La boîte de dialogue de la figure 6.20 apparaît. Faites passer les variables de gauche à droite à l'aide de la flèche, puis cliquez sur **OK.**

Les résultats de la corrélation apparaissent dans le tableau de résultats à la figure 6.21.

Le tableau des corrélations obtenu entre la note du trimestre, l'absentéisme en nombre de jours par trimestre et la qualité perçue de l'enseignement montre qu'il existe un lien significatif entre l'absentéisme et la note moyenne. Le signe – atteste que ces deux variables évoluent de manière inversement proportionnelle. Autrement dit, plus un étudiant est absent moins sa note du trimestre est bonne. Il n'y a pas de corrélation entre la note moyenne du trimestre et la qualité perçue de l'enseignement, comme il n'y a pas de lien entre le taux d'absentéisme des étudiants et la qualité perçue de l'enseignement.

Figure 6.19

Rappel de la procédure pour rentrer les données sous SPSS *(suite)*.

Fichier Edition Affichage Données Transformer A

1 : moyenne 5

	moyenne	qualité	absentéism
1	5	4	10
2	7	4	2
3	15	3	0
4	11	2	0
5	16	3	0
6	12	2	1
7	11	3	0
8	14	4	0
9	10	4	0
10	14	4	0
11	11	3	0
12	9	4	1
13	9	2	2
14	11	4	0
15	10	4	0
16	7	2	1
17	14	4	0
18	15	4	0
19	11	3	0
20	14	4	0
21	12	3	0
22	11	4	0
23	9	3	0
24	8	2	2
25	11	4	0
26	10	3	0
27	14	4	0
28	12	4	0

Figure 6.20

Demande de corrélation entre la qualité de l'enseignement, l'absentéisme et la moyenne.

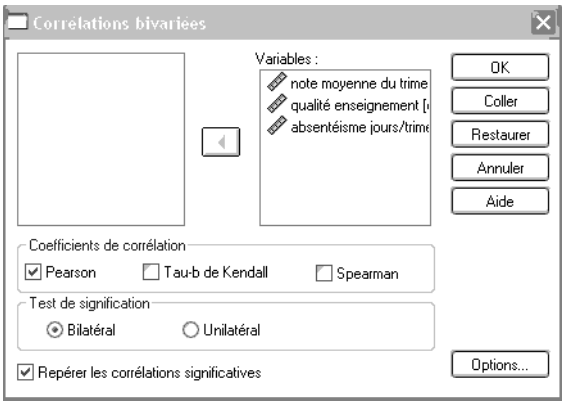


Figure 6.21

Résultats de la corrélation entre la moyenne, l'absentéisme et la qualité de l'enseignement.

Corrélations

		note moyenne du trimestre	absentéisme jours/trimest	qualité enseignement
note moyenne du trimestre	Corrélation de Pearson	1	-,611**	,215
	Sig. (bilatérale)		,001	,271
	N	28	28	28
absentéisme jours/trimest	Corrélation de Pearson	-,611**	1	,005
	Sig. (bilatérale)	,001		,979
	N	28	28	28
qualité enseignement	Corrélation de Pearson	,215	,005	1
	Sig. (bilatérale)	,271	,979	
	N	28	28	28

**La corrélation est significative au niveau 0.01 (bilatéral).

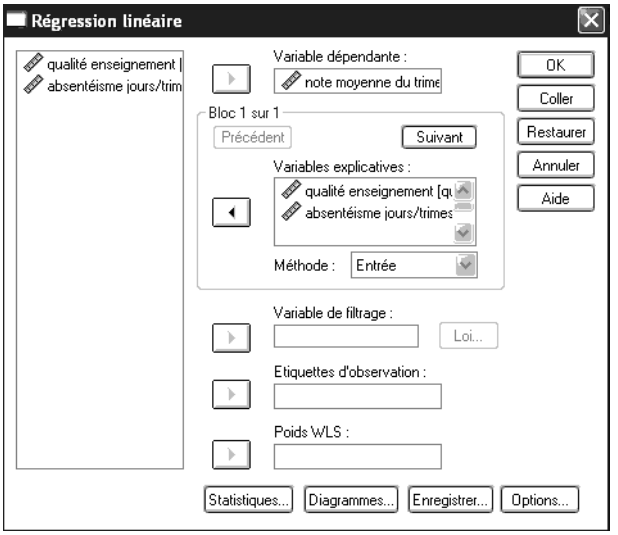
3. Pour réaliser la régression multiple (pas à pas), allez dans le menu **Analyse** > **Régression** > **Linéaire**.

Transférez la variable note moyenne vers **Variable dépendante**, puis les variables qualité de l'enseignement et absentéisme, chacune à leur tour, vers **Variables explicatives** en cliquant sur les flèches.

La méthode de sélection des variables par défaut est choisie. Cliquez ensuite sur **OK** (voir figure 6.22).

Figure 6.22

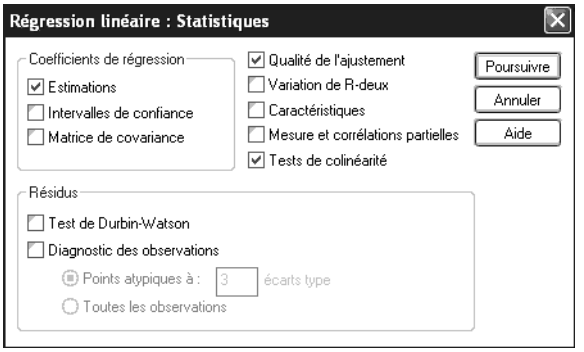
Demande d'une régression multiple (Pas à pas) sur la moyenne du trimestre.



Avant de lancer la commande de régression, la non-colinéarité entre les variables indépendantes doit être vérifiée. Pour ceci, il faut faire des tests de colinéarité. Retournez dans la boîte de dialogue puis cliquez sur **Statistiques, Tests de colinéarité**. Cliquez ensuite sur **Poursuivre** et **OK** (voir figure 6.23).

Figure 6.23

Demande de tests de colinéarité.



Les résultats de la régression multiple figurent dans la boîte de résultats à la figure 6.24.

Figure 6.24

Interprétation des résultats de la régression multiple sur la moyenne du trimestre.

Variables introduites/éliminées[b]

Modèle	Variables introduites	Variables éliminées	Méthode
1	absentéisme jours/trimest, qualité enseignement[a]	.	Introduire

a. Toutes variables requises introduites

b. Variable dépendante : note moyenne du trimestre

Toutes les variables sont conservées pour la régression puisque nous n'avons pas spécifié de méthode de régression particulière ; la méthode par défaut prend toutes les variables explicatives (voir figure 6.25).

Figure 6.25

Interprétation des résultats de la régression multiple sur la moyenne du trimestre *(suite)*.

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,648[a]	,420	,374	2,132

a. Valeurs prédites : (constantes), absentéisme jours/trimest, qualité enseignement

ANOVA[b]

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	82,450	2	41,225	9,068	,001[a]
	Résidu	113,657	25	4,546		
	Total	196,107	27			

a. Valeurs prédites : (constantes), absentéisme jours/trimest, qualité enseignement

b. Variable dépendante : note moyenne du trimestre

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Statistiques de colinéarité	
		B	Erreur standard	Bêta	t	Signification	Tolérance	VIF
1	(constante)	9,221	1,816		5,076	,000		
	qualité enseignement	,755	,526	,218	1,434	,164	1,000	1,000
	absentéisme jours/trimest	-,848	,211	-,612	-4,017	,000	1,000	1,000

a. Variable dépendante : note moyenne du trimestre

Les deux variables prises en compte expliquent 37 % de la moyenne du trimestre (R^2 ajusté). Le tableau ANOVA atteste que le modèle est significatif (Signification = 0,00).

Tolérances et facteurs d'inflation de la variance (VIF) sont égaux à 1, ce qui montre que les variables explicatives sont peu corrélées entre elles et représentent un indice de qualité du modèle.

Le test *t* de la régression de la qualité de l'enseignement sur la note moyenne du trimestre n'est pas significatif ($p > 0,0$; $t < 2$), nous pouvons donc conclure que seul l'absentéisme a un effet significatif et négatif sur la moyenne du trimestre ($t = -4,01$; $p < 0,01$). Autrement dit, plus les étudiants sont absents, moins leur moyenne est bonne.

EXERCICE 3 ÉTUDE DU POINT DE VENTE

Énoncé

Vous travaillez sur une enquête destinée à mieux comprendre les comportements d'achat de clients d'un magasin de chaussures. Vous cherchez à identifier ces clients et connaître leur attitude à l'égard du point de vente. Vous avez collecté 400 réponses et cherchez à exploiter ces données (fichier « pointdevente » [1]).

Le gérant du magasin souhaite savoir :

1. s'il existe une relation entre la taille du foyer et le montant dépensé dans le magasin ;
2. si le niveau d'études influence l'intention d'achat dans le point de vente.

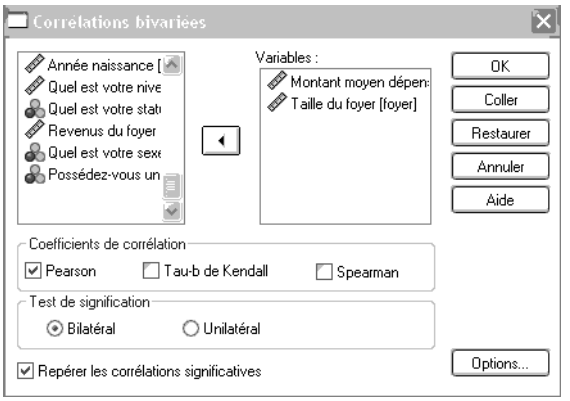
Solution

1. Une corrélation permet de savoir si la taille du foyer et le montant dépensé dans le magasin sont liés.

Voici la commande à effectuer : **Analyse** > **Corrélation** > **Bivariée**. Faites passer les variables foyer et montant vers **Variables**, puis **OK** (voir figure 6.26).

Figure 6.26

Demande de corrélation entre la taille du foyert et le montant dépensé.



Les résultats de la corrélation figurent dans la boîte de résultats à la figure 6.27.

Figure 6.27

Interprétation des résultats de la corrélation entre la taille du foyert et le montant dépensé.

Corrélations

		Montant moyen dépensé par mois	Taille du foyer
Montant moyen dépensé par mois	Corrélation de Pearson	1	,038
	Sig. (bilatérale)		,445
	N	400	400
Taille du foyer	Corrélation de Pearson	,038	1
	Sig. (bilatérale)	,445	
	N	400	400

L'absence de signe ** indique que la corrélation entre la taille du foyer et le montant dépensé dans le magasin n'est pas significative. Il n'y a pas de lien entre ces deux variables.

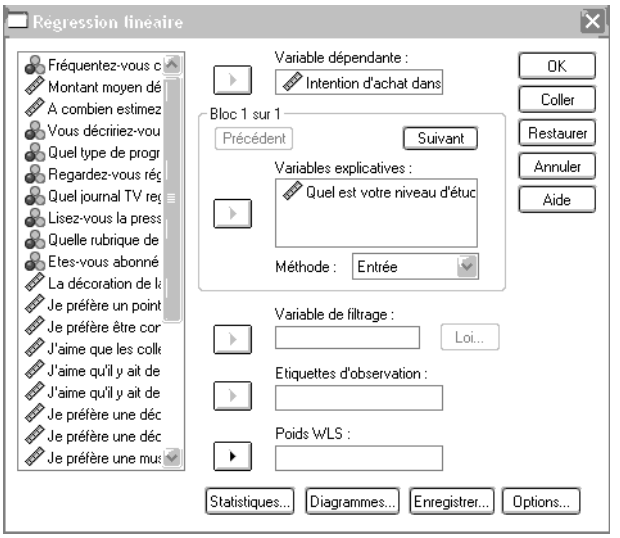
1. Vous trouverez ce fichier à l'adresse : http://www.pearsoneducation.fr.

2. Pour savoir si le niveau d'études influence l'intention d'achat au point de vente, il faut faire une régression du niveau d'études sur l'intention d'achat.

Voici la commande à effectuer : **Analyse** > **Régression** > **Linéaire**. Faites passer l'intention d'achat vers **Variable dépendante** et le niveau d'études (Question : Quel est votre niveau d'études ?) vers **Variables explicatives**, puis faites **OK** (voir figure 6.28).

Figure 6.28

Demande de régression du niveau d'études sur l'intention d'achat.



Les résultats de la régression apparaissent dans les tableaux de la figure 6.29.

Figure 6.29

Interprétation des résultats de la régression du niveau d'études sur l'intention d'achat.

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,502[a]	,252	,250	1,131

a. Valeurs prédites : (constantes), Quel est votre niveau d'étude?

ANOVA[b]

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	171,551	1	171,551	134,155	,000[a]
	Résidu	508,946	398	1,279		
	Total	680,498	399			

a. Valeurs prédites : (constantes), Quel est votre niveau d'étude?

b. Variable dépendante : Intention d'achat dans le point de vente

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification	Statistiques de colinéarité	
		B	Erreur standard	Bêta			Tolérance	VIF
1	(constante)	,448	,238		1,882	,061		
	Quel est votre niveau d'étude?	,464	,040	,502	11,583	,000	1,000	1,000

a. Variable dépendante : Intention d'achat dans le point de vente

Le tableau ANOVA atteste que le modèle est significatif. Le test *t* de la régression du niveau d'études sur l'intention d'achat est supérieur à 2. Nous pouvons donc conclure que le niveau d'études a un effet significatif et positif sur l'intention d'achat dans le magasin ($t = 11{,}58$; $p < 0{,}01$). Le niveau d'études explique 25 % de l'intention d'achat (R^2 ajusté). Autrement dit, plus les individus ont des diplômes, plus ils sont prêts à effectuer des achats dans le magasin.

Chapitre 7

L'analyse conjointe

L'analyse conjointe est une forme d'analyse de variance qui permet de mesurer les préférences des individus relatives aux attributs d'un objet. L'objectif est d'identifier les préférences globales et de mesurer l'effet conjoint de caractéristiques. Par exemple, il s'agira de cerner les critères les plus importants dans l'achat d'un ordinateur (puissance, marque, design, etc.). Cette technique permet d'expliquer les préférences pour un objet en fonction de ses caractéristiques, de déduire l'importance de ces différentes caractéristiques et de leurs modalités dans l'évaluation globale portée par les individus. L'analyse conjointe est surtout utilisée dans le contexte des biens de consommation, où elle sert à améliorer les produits ou les services selon les résultats.

Nous verrons dans un premier temps les principes de l'analyse conjointe avant de présenter le déroulement, la réalisation et l'interprétation de cette méthode.

1 Les principes de l'analyse conjointe

Les travaux de Green dans les années 1970 marquent le début de la prise en compte de l'analyse conjointe dans la recherche en marketing. La méthode dite de l'« analyse des mesures conjointes », au développement croissant depuis les années 1980, vise à mieux comprendre le comportement des individus et, en particulier, du consommateur.

Le sketch de Coluche illustre la problématique de l'analyse conjointe sur la préférence entre être « *grand, riche, beau et intelligent* » et « *petit, pauvre, moche et bête* ». Si l'on présente les combinaisons suivantes « *grand, riche, moche et bête* » ou « *petit, pauvre, beau et intelligent* », l'individu est alors amené à faire des compromis dans lesquels l'avantage d'une caractéristique compense une autre qu'il doit rejeter. L'analyse conjointe permet de répondre aux questions suivantes : Quelle est l'importance de tel ou tel attribut (prix, dimensions, etc.) du produit pour le consommateur ? Quelle est l'importance de certains niveaux d'attributs (niveau de prix, dimensions en cm, etc.) par rapport à d'autres dans l'esprit du consommateur ? Cette méthode détermine à la fois l'importance relative de chaque attribut et les niveaux des attributs préférés des répondants.

Lorsqu'on dispose d'informations sur les répondants (données démographiques ou autres), l'analyse conjointe permet d'identifier les segments de marché pour lesquels des produits spécifiques seront plus adaptés. Par exemple, une personne appartenant à une CSP+ et un étudiant peuvent avoir des goûts différents auxquels des offres de produits distincts pourront répondre.

L'analyse conjointe repose sur la décomposition de la préférence en utilités partielles. Pour déterminer l'utilité totale d'un produit, on suppose que l'individu additionne les utilités partielles des attributs du produit. On parle de *modèle additif*. Au final, l'individu choisit parmi les produits celui qui lui procure l'utilité totale la plus élevée. L'estimation permet ainsi d'obtenir, pour chaque facteur et ses niveaux, des utilités partielles ainsi que l'importance de chaque attribut. Ce qui compte est donc l'individu tel qu'il réagit dans une situation déterminée.

L'analyse conjointe appartient aux modèles de décomposition (voir focus 7.1 sur le modèle compensatoire) où l'importance des caractéristiques est estimée à partir des préférences déclarées du consommateur et de ses notations des différents produits sur plusieurs caractéristiques. Elle permet d'analyser l'importance des caractéristiques du produit dans la formation des préférences.

Focus 7.1 Le modèle compensatoire

Le modèle d'attitude implicite de l'analyse conjointe est un modèle compensatoire, où l'évaluation se fonde sur le principe du compromis, c'est-à-dire qu'un peu moins d'un attribut peut être compensé par un peu plus d'un autre.

Par exemple, un individu qui cherche un appartement peut avoir plusieurs critères : le montant du loyer, la superficie, le nombre de pièces, la luminosité, la proximité des transports publics, etc. Si l'un de ces critères n'est pas satisfait (superficie insuffisante), il peut être compensé par un montant du loyer plus faible dans un modèle compensatoire (ce n'est pas le cas dans le modèle non compensatoire).

1.1 Les utilisations de l'analyse conjointe en marketing

L'analyse conjointe est largement utilisée en marketing pour l'identification d'un nouveau concept, pour divers tests (prix, produits, publicité, distribution, etc.), pour l'analyse concurrentielle ou la segmentation du marché (voir tableau 7.1). Il s'agit, par exemple :

- de déterminer l'importance relative d'attributs dans le processus de choix des consommateurs ;
- d'estimer la part de marché des marques qui diffèrent au niveau des attributs ;
- de déterminer la composition d'objets les plus appréciés ;
- de segmenter le marché à partir des similarités de préférences pour des niveaux d'attributs.

Tableau 7.1 : Les applications de l'analyse conjointe

Pour les biens de consommation courante	
Nouveaux produits	72 %
Prix	61 %
Segmentation	48 %
Publicité	39 %
Distribution	7 %

1.2 Les conditions d'application de l'analyse conjointe

L'analyse des mesures conjointes nécessite que **les variables explicatives soient qualitatives ou nominales et que les variables à expliquer soient quantitatives**. Ces dernières peuvent être évaluées :

- à l'aide d'une échelle de mesure ;
- à l'aide d'un ordre de préférence (classement) entre différentes combinaisons de niveaux de facteurs.

Par exemple, dans le cadre d'une étude sur la préférence des clients à l'égard d'une offre de transport aérien, on pourra demander aux individus d'évaluer différentes offres de compagnies avec une note de 1 à 9, autrement dit de les classer. Ces offres incluront, par exemple, le prix du billet (bas, moyen, élevé), la qualité du service à bord (excellente, moyenne, mauvaise), le nombre d'escales, etc. Il sera ainsi possible d'estimer, parmi ces facteurs et leurs niveaux, celui qui a le plus d'importance pour les clients dans leur choix d'une offre de transport aérien. L'objectif est ensuite d'élaborer une offre optimale pour la clientèle.

La méthodologie de l'analyse conjointe est jalonnée par trois grandes étapes (Green et Srinivasan, 1990) :

- **collecte des données** : choix du plan factoriel complet ou fractionné, de la forme des questions et de la méthode de recueil ;
- **définition de l'échelle de mesure de la variable dépendante** : choix de la mesure (classement, notation des combinaisons, comparaison de paires de combinaisons) ;
- **estimation** : étape liée à la nature de la mesure de la variable dépendante (ANOVA si la variable dépendante est quantitative, analyse monotone de la variance si elle est ordinale).

Ces étapes sont développées au cours de la section suivante.

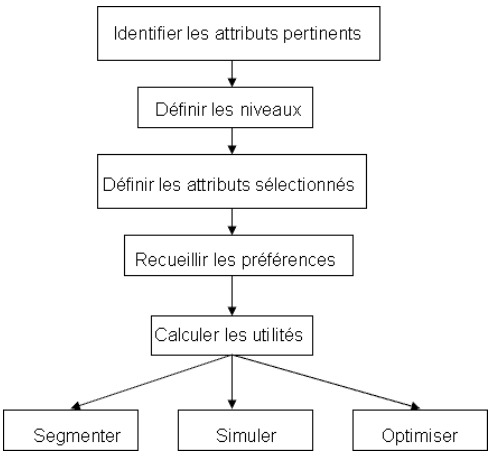
2 La préparation de l'analyse conjointe

L'analyse conjointe demande au préalable la résolution d'un certain nombre de questions. En effet, avant même de collecter les données, le chargé d'étude doit s'interroger sur les attributs qu'il cherche à évaluer et leurs niveaux. Cette première sélection des attributs ou facteurs les plus importants et de leurs niveaux ou modalités déterminera le plan d'expérience et le mode de collecte des données.

Le schéma de la figure 7.1 présente les différentes phases de la méthode de l'analyse conjointe et met l'accent sur la première étape de la sélection des attributs et des niveaux.

Figure 7.1

Les étapes de l'analyse conjointe.



Source : adapté de Liquet et Benavent.

2.1 La sélection des attributs ou facteurs et de leurs niveaux

La sélection des variables et des niveaux à tester est cruciale. Les facteurs choisis doivent correspondre à l'ensemble des critères intervenant dans le choix des individus. Il est donc primordial que tous les attributs soient importants et indépendants, qu'ils décrivent complètement le produit et qu'ils soient manipulables. Par exemple, pour identifier le packaging de biscuits pour le goûter des enfants, le chargé d'étude sélectionnera le format du paquet, le type d'emballage, le code couleurs, etc., et en termes de niveaux, deux tailles pour le format (moyen, petit), deux pour le type d'emballage (carton, sachet) et trois pour le code couleurs (bleu-rouge ; bleu-vert ; bleu-jaune).

Pour synthétiser, les attributs ou facteurs doivent être :

- **Exhaustifs.** Il ne doit pas manquer de facteurs majeurs dans l'évaluation de l'objet.
- **Importants.** Les facteurs sélectionnés doivent être essentiels dans l'évaluation de l'objet par les individus.
- **Indépendants.** Les facteurs ne doivent pas être corrélés, sinon c'est la redondance qui est mesurée au lieu de la préférence.

Si plusieurs attributs sont fortement corrélés, il faut alors essayer de les regrouper en un facteur globalisant. À titre d'exemple, voici quatre attributs employés pour décrire l'ambiance d'un restaurant : le décor (raffiné ou simple), l'agencement (banal ou élaboré), la lumière (tamisée ou non) et le niveau sonore (élevé, moyen, faible). Ces attributs pourront être regroupés sous une variable unique, intitulée « ambiance du restaurant » et comptant trois modalités (agréable, neutre ou désagréable). Les autres facteurs pourront être le tarif, l'emplacement du restaurant, la variété des plats, la qualité du service, etc.

Les attributs et leurs niveaux doivent être importants et proches du réel des individus qui seront interrogés. Ils doivent aussi être suffisamment contrastés afin que les personnes puissent se prononcer. Enfin, le panier d'attributs doit être manipulable lors d'une simulation. Reprenons l'exemple de l'emballage de biscuits : le format, le type d'emballage et le code couleurs sont indépendants. À propos de leur importance, le chercheur doit s'assurer au préalable, par exemple grâce à une étude qualitative, que les attributs clés sont sélectionnés. Des entretiens avec des enfants sur leurs préférences concernant l'emballage de biscuits pour le goûter serviront à identifier les facteurs et leurs modalités clés.

Il est généralement recommandé d'avoir un nombre de niveaux équilibré, c'est-à-dire équivalent pour chaque attribut. Cela permet d'éviter que les individus accordent plus de poids aux attributs présentant davantage de niveaux que les autres. Un nombre limité de niveaux (2 ou 3) permet, en outre, de présenter aux personnes interrogées des options plus claires. Dans le cas de notre exemple, le facteur code couleurs a trois niveaux, ce qui peut lui attribuer une plus grande importance. Il faut en tenir compte dans l'analyse.

Focus 7.2 Comment identifier les attributs déterminants ?

Afin d'optimiser la phase clé de sélection des attributs et de leurs niveaux, une étude exploratoire est souvent nécessaire. Elle vise à repérer quels sont, pour les individus, les attributs les plus importants et leurs niveaux. Elle permet de s'assurer que les facteurs auxquels le chargé d'étude a pensé sont pertinents et qu'il n'a pas oublié de facteurs ou niveaux importants.

Nous prenons ici l'exemple d'une étude sur les préférences des consommateurs à l'égard de crèmes glacées. Un questionnaire proposé à 300 ménages a permis d'identifier les attributs

déterminants du processus d'achat de crèmes glacées. Ils ont dû évaluer chacun de ces critères selon le degré d'importance (de *« pas du tout »* à *« très important »*) :

- le prix ;
- le parfum ;
- la marque ;
- le conditionnement.

Une question ouverte (Autre) a permis d'identifier d'autres critères clés, par exemple :

- la composition du produit.

La sélection du nombre de facteurs a une incidence directe sur le plan d'expérience utilisé. En effet, lorsqu'on cherche à tester un nombre important d'attributs et de niveaux, il est souvent difficile de considérer toutes les configurations possibles (voir focus 7.3, Plan complet et plan fractionné) car c'est trop coûteux en termes d'expérience.

Focus 7.3 Plan complet et plan fractionné

Lorsqu'on utilise un plan complet, chaque profil décrit un objet complet, c'est-à-dire une combinaison différente de niveaux de facteurs pour tous les facteurs ou tous les attributs. Cette méthode permet donc d'avoir une évaluation de toutes les combinaisons possibles.

Cependant, lorsqu'on cherche à évaluer beaucoup d'attributs et de niveaux, le nombre total de profils à évaluer devient trop important pour que les répondants puissent les classer ou les noter de façon cohérente. Il est alors préférable d'utiliser un plan fractionné qui présente une fraction appropriée de toutes les combinaisons possibles de niveaux de facteurs.

L'ensemble qui en résulte, appelé « tableau orthogonal », est conçu pour saisir les effets principaux de chaque niveau de facteur.

2.2 La mise en œuvre de la simulation

La construction de scénarios (voir focus sur la méthode des scénarios vue au chapitre 6) détermine la crédibilité de l'expérience. En effet, l'analyse conjointe repose sur des informations collectées auprès d'individus à qui l'on demande de faire des choix, de déclarer leurs préférences entre divers objets présentés avec des *stimuli*. Les *stimuli* sont des combinaisons d'attributs et de leurs niveaux qui sont évalués en fonction de leurs attraits. L'analyse conjointe repose sur la méthode des scénarios.

La méthode des scénarios ou des vignettes

La méthode dite des scénarios ou des vignettes s'appuie sur l'utilisation de scénarios. Le scénario simule de vraies expériences, comme celle du choix d'un paquet de biscuits pour le goûter. Les individus, mis dans une situation hypothétique, doivent répondre « comme si » ils se trouvaient réellement dans ces situations. En prenant l'exemple du choix du paquet de biscuits, on demandera aux enfants de choisir entre les différents emballages qui pourront être dessinés. Ils choisiront, non pas selon des questions directement posées sur leurs préférences en termes de format (petit ou grand), de code couleurs, etc., mais en fonction de combinaisons de niveaux d'attributs.

La mise en situation possède l'avantage de rendre l'expérience plus réaliste et, par là, de mieux impliquer les répondants dans la création de sens. Le but est aussi de s'éloigner de la rationalisation et d'effectuer son choix comme dans une situation réelle d'achat.

Une description verbale, un texte descriptif écrit, une photo, un dessin ou un prototype peuvent servir à présenter les *stimuli*. Il faut ici veiller à ce que les scénarios soient crédibles et qu'aucun ne soit manifestement trop attractif ou, au contraire, répulsif. Les *stimuli* doivent, de fait, avoir une apparence similaire afin que les préférences des individus soient bien le fruit des attributs testés.

Le tableau 7.2 expose les méthodes de présentation des scénarios les plus usitées.

Tableau 7.2 : Les méthodes de présentation des *stimuli* les plus utilisées de l'analyse conjointe

Description verbale	50 %
Texte descriptif	20 %
Descriptif visuel	19 %
Prototype	7 %
Autres	4 %

Source : Cattin et Wittink, 1982.

Exemple de vignettes

Voici un exemple de vignettes utilisées pour décrire différentes offres de chambres d'hôtel. Les facteurs manipulés sont :

- la localisation de l'hôtel (centre-ville, proche d'une autoroute, excentré) ;
- le type de chambre (luxe, standard, simple) ;
- le prix de la chambre (élevé, économique) ;
- la marque (chaîne d'hôtel connue ou pas).

Le tableau 7.3 liste les vignettes qui permettent de décrire les diverses combinaisons présentées aux individus.

Tableau 7.3 : Vignettes décrivant les offres de chambres d'hôtel

Description des concepts	Évaluation Classement ou Note
1. Chambre standard de la chaîne d'hôtels connue Y excentré, économique	
2. Chambre de luxe de la chaîne d'hôtels connue Y excentré, économique	
3. Chambre de luxe de la chaîne d'hôtels connue Y centre-ville, prix élevé	

Tableau 7.3 : Vignettes décrivant les offres de chambres d'hôtel *(suite)*

Description des concepts	Évaluation Classement ou Note
4. Chambre simple de la chaîne d'hôtels connue Y proche d'une autoroute, économique	
5. Chambre simple, proche d'une autoroute, économique	
6. Chambre de luxe, centre-ville, prix élevé	
…	

2.3 La collecte des données

Lors de la phase de collecte des données, le chargé d'étude s'interroge sur le choix de la méthode de comparaison des vignettes ou des scénarios à tester, du mode d'administration et du mode de recueil des données. Nous verrons successivement ces trois points clés.

Le choix de la méthode de comparaison des scénarios

Trois méthodes sont le plus souvent utilisées pour la collecte des données mais c'est celle du profil complet qui est la plus courante. Avec **la méthode du profil complet**, chacun des répondants est exposé à toutes les combinaisons de niveaux de facteurs. Elles sont décrites séparément et l'individu doit évaluer chacune d'elles. Si on reprend l'exemple du choix d'une chambre d'hôtel, les individus devraient noter les différentes offres décrites par les vignettes, soit un total de 36.

Par conséquent, il est recommandé de sélectionner un nombre de facteurs inférieur à 6 et un nombre de niveaux pour chacun de ces facteurs limité à 3 ou 4. En effet, même si l'utilisation de plans fractionnés (voir chapitre 5) permet de réduire les profils que les individus évaluent, il existe des biais de réponses dus à un effet d'apprentissage. L'ordre de présentation des concepts influence l'appréciation des répondants, il faut donc veiller à présenter les vignettes de façon aléatoire.

La méthode de comparaison par paires présente les différentes vignettes ou les différents scénarios deux à deux. Les personnes interrogées estiment les paires de concepts jusqu'à ce qu'ils soient tous évalués. Si on reprend l'exemple de l'emballage de biscuits pour le goûter, on demandera aux individus de choisir parmi les propositions (combinaison x de niveaux de facteurs) : un sachet de petit format et un code couleurs bleu-rouge seront comparés à un carton de petit format et un code couleurs bleu-jaune… Cela jusqu'à ce qu'ils aient évalué toutes les combinaisons. Il est aussi possible de faire noter différentes paires à des groupes d'individus différents, ce qui permet de gagner du temps.

La méthode du trade off utilise deux facteurs à la fois. On demande aux répondants de classer par ordre de préférence toutes les combinaisons des niveaux des deux facteurs. Cette méthode repose sur une matrice qui croise l'ensemble de combinaisons des niveaux de facteurs deux à deux. Elle a pour inconvénient majeur sa lourdeur lorsque le nombre de facteurs est élevé.

L'avantage des méthodes du trade off et de comparaison par paires est que les individus perçoivent clairement les offres et répondent facilement. Leur principal inconvénient est le nombre important d'évaluations nécessaires. La méthode du profil complet en exige moins. Elle est donc intéressante lorsqu'on cherche à tester un nombre important d'attributs et de niveaux d'attributs.

Pour les deux méthodes, il n'est pas toujours utile de faire évaluer toutes les combinaisons possibles même si l'emploi de plans fractionnés (voir focus 7.3) peut parfois pallier ce problème.

Le choix du mode d'administration de l'enquête

L'information collectée est déduite des évaluations d'objets par les individus interrogés. La taille de l'échantillon, issu de la population cible de l'objet testé, varie entre 100 et 1 000. Elle doit être suffisamment importante pour assurer la fiabilité des résultats.

Une fois l'échantillon choisi, le chargé d'étude propose les scénarios ou profils à chaque répondant. Selon le nombre de scénarios à tester, il peut administrer soit la totalité des configurations possibles (plan complet), soit une sélection (plan fractionné).

Les données sont ensuite collectées (voir tableau 7.4), le plus souvent *via* des entretiens de groupe, des réunions, en interrogeant les individus directement ou encore en suivant un protocole. Il est recommandé de travailler avec un expert du domaine (chef de produit, par exemple) pour élaborer des scénarios réalistes et diffuser l'étude. Ces interventions permettent de renforcer la validité de l'étude.

Tableau 7.4 : Les conditions d'administration de l'analyse conjointe

Interventions d'experts dans l'étude (manager, etc.)	30 %
Entretiens de groupe	26 %
Questionnements directs d'individus	18 %
Autres	26 %

Le choix du mode de recueil des données

Dans l'analyse conjointe, la variable mesurée est généralement la préférence ou l'intention d'achat. Les individus interrogés fournissent donc un score ou un rang en fonction de leurs goûts et de leurs intentions d'achat.

Trois méthodes de recueil des données existent. On peut demander aux individus :

- de donner **un score à chaque profil**, selon leurs préférences (donnée métrique) ;
- d'assigner **un rang à chaque profil**, de 1 jusqu'au nombre total de profils ;
- de **trier les profils en termes de préférences** par ordre croissant ou décroissant (classement).

Les données sont le plus souvent recueillies à partir d'un classement des profils ou à l'aide d'un score sur chaque profil.

Certains chercheurs estiment que le classement ou l'attribution d'un rang reflète plus précisément le comportement des individus lors de la prise d'une décision. D'autres, tenants des données d'évaluation, pensent, au contraire, qu'elles sont plus pratiques pour les répondants.

Dans les deux cas, étant donné que l'ordre de présentation des *stimuli* peut affecter l'évaluation des répondants, il est recommandé de varier l'ordre de présentation des scénarios.

3 L'interprétation de l'analyse

C'est à partir d'une application que nous verrons comment réaliser et interpréter une analyse conjointe. Dans un premier temps, nous voyons la sélection des attributs, l'enregistrement des réponses, la création des scénarios et du plan *via* le logiciel SPSS. Ensuite, nous présentons les résultats de l'analyse conjointe ainsi que les procédures de vérification de sa fiabilité et de validité.

3.1 Exemple de réalisation et d'interprétation de l'analyse conjointe

Le choix des attributs et de leurs niveaux

Notre étude porte sur le challenge de vente préféré des commerciaux. Il ressort d'entretiens avec eux que trois caractéristiques des challenges sont très importantes : le type d'objectif, le format de compétition et le budget (fermé, avec un nombre de gagnants et un budget défini à l'avance, ou ouvert, avec un quota à atteindre et un nombre de gagnants et un budget méconnu au départ). Ces trois caractéristiques indépendantes les unes des autres sont donc sélectionnées (condition essentielle de l'analyse conjointe).

Le nombre de niveaux d'attributs est équilibré à deux : l'objectif est soit quantitatif (chiffre d'affaires), soit qualitatif (évaluation de la connaissance des produits), le format est soit individuel (objectif assigné individuellement), soit en équipe (objectif collectif) et le budget est soit ouvert (niveau à atteindre précis), soit fermé (classement). Trois caractéristiques de base, ayant chacune deux modalités sont sélectionnées (voir tableau 7.5).

Tableau 7.5 : Les attributs sélectionnés et leurs niveaux

Format de compétition	Individuel
	En équipe
Type d'objectif	Quantitatif
	Qualitatif
Budget	Ouvert
	Fermé

Le nombre d'attributs et de niveaux étant faible, nous pouvons utiliser la méthode du plan complet. Nous devrions avoir : $2 \times 2 \times 2 = 8$ profils.

Nous allons voir maintenant la procédure à suivre sur SPSS pour générer le plan qui servira de base au développement des scénarios de l'analyse conjointe.

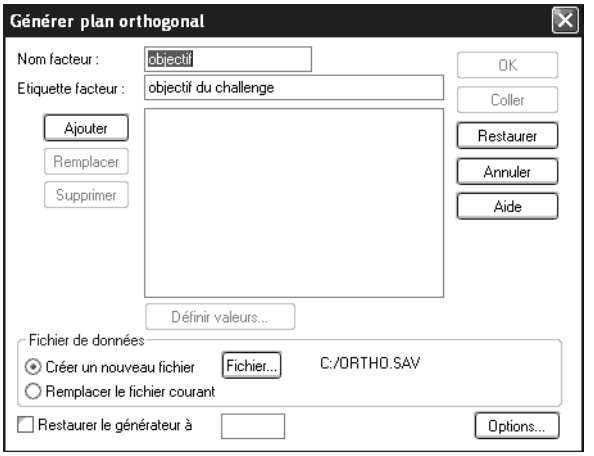
La création du plan orthogonal

Pour créer le plan orthogonal, allez dans le menu **Données** > **Plan orthogonal** > **Générer**. Dans la boîte de dialogue qui apparaît (voir figure 7.2), saisissez le nom et l'étiquette de chacun des attributs ou facteurs.

Ici, nous commençons par inscrire objectif dans le champ **Nom facteur** et objectif du challenge dans le champ **Etiquette facteur.**

Figure 7.2

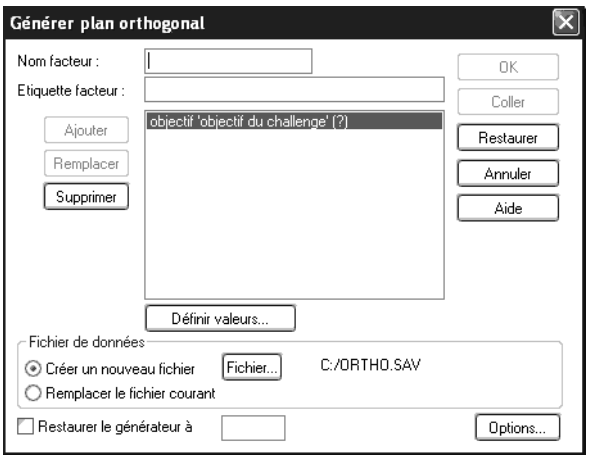
Création du plan orthogonal.



Cliquez ensuite sur le bouton **Ajouter** pour insérer les autres facteurs, toujours dans les champs **Nom facteur** et **Etiquette facteur**. Nous saisissons format et format du challenge puis, après avoir cliqué sur **Ajouter :** budget et budget du challenge.

Figure 7.3

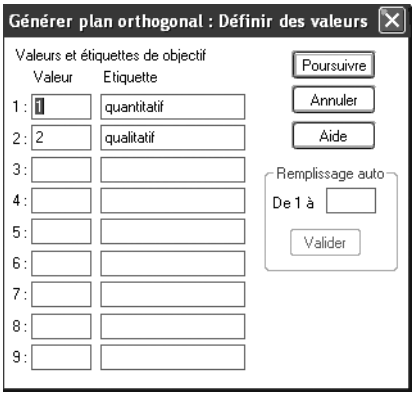
Enregistrement des attributs.



Ensuite, il faut définir les niveaux de chacun des facteurs. Pour cela, sélectionnez chaque facteur et cliquez sur le bouton **Définir valeurs** (voir figure 7.3). Nous choisissons d'attribuer la valeur 1 au format individuel et la valeur 2 au format en équipe. Cliquez ensuite sur **Poursuivre** pour revenir à la boîte de dialogue et refaites cette action pour chacun des attributs (la valeur 1 pour les objectifs quantitatifs et la valeur 2 pour les objectifs qualitatifs ; idem pour le budget du challenge, la valeur 1 pour le budget fermé et la valeur 2 pour le budget ouvert) [voir figure 7.4].

Figure 7.4

Enregistrement des niveaux des attributs.



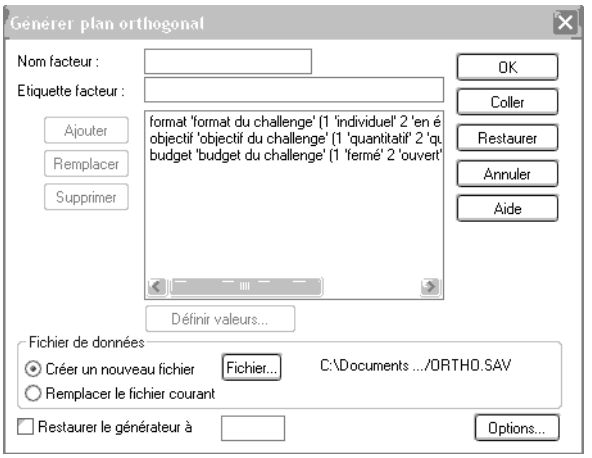
Ensuite, vous devez indiquer où placer ce plan orthogonal. Pour cela, cliquez sur le bouton **Fichier** de la boîte de dialogue visible à la figure 7.5 (choisissez un emplacement où il peut être facilement récupéré et souvenez-vous de son intitulé).

Attention ! Le nom du fichier doit être en majuscules sinon SPSS ne génère pas le plan orthogonal.

Nous appelons le fichier ORTHO7.

Figure 7.5

Sauvegarde du fichier et demande de copie de la syntaxe.



La procédure de l'analyse des mesures conjointes sous le logiciel SPSS requiert un mini-programme ou une macrocommande qui indique au logiciel les facteurs étudiés ainsi que leurs niveaux. SPSS crée ainsi le plan factoriel qui donne les combinaisons à tester (ORTHO).

Ensuite, le logiciel croise le plan factoriel avec les réponses enregistrées dans le fichier (DATA) comme nous allons le découvrir. Très important pour la suite de l'analyse : cliquez sur le bouton **Coller** de la même boîte de dialogue afin d'avoir la syntaxe de la macrocommande (voir figure 7.6).

C'est seulement après cette action que vous cliquez sur **OK** (pour cela, revenez au menu **Données** > **Plan orthogonal** > **Générer**). Le message suivant (voir figure 7.7) doit alors apparaître dans **Résultats**.

Figure 7.6

Copie de la syntaxe.

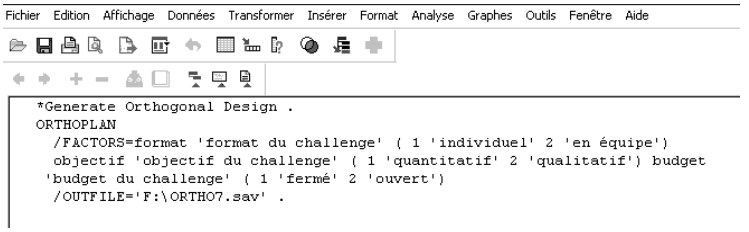


Figure 7.7

Annonce de la création du plan orthogonal.

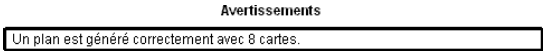
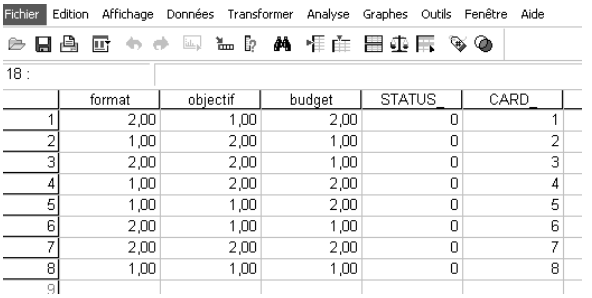


Figure 7.8

Affichage du plan orthogonal.



	format	objectif	budget	STATUS_	CARD_
1	2,00	1,00	2,00	0	1
2	1,00	2,00	1,00	0	2
3	2,00	2,00	1,00	0	3
4	1,00	2,00	2,00	0	4
5	1,00	1,00	2,00	0	5
6	2,00	1,00	1,00	0	6
7	2,00	2,00	2,00	0	7
8	1,00	1,00	1,00	0	8
9					

À ce stade, le plan généré, qui se trouve à l'emplacement choisi, permet de créer les scénarios. Le scénario n° 2 présentera aux individus un challenge individuel (format = 1), un objectif quantitatif (format = 2) et un budget fermé (budget = 2).

La création des scénarios

Afin de rendre l'expérience plus réaliste, nous optons pour une présentation des profils de challenges à travers une simulation destinée à mettre les commerciaux en situation pour effectuer leur arbitrage. Nous créons, avec l'aide de managers, des scénarios de challenges ayant déjà été utilisés au sein de différentes entreprises. Les personnes interrogées en connaissent le principe, ce qui renforce le réalisme de la simulation.

Après une brève introduction sur la situation du vendeur au sein de l'entreprise, on annonce que la direction cherche à connaître le challenge préféré des commerciaux. Il leur est demandé d'en classer huit par ordre décroissant de préférence sur une grille de classement.

Voici un exemple du scénario n° 8, qui correspond à la dernière ligne du plan sous SPSS (voir figure 7.9).

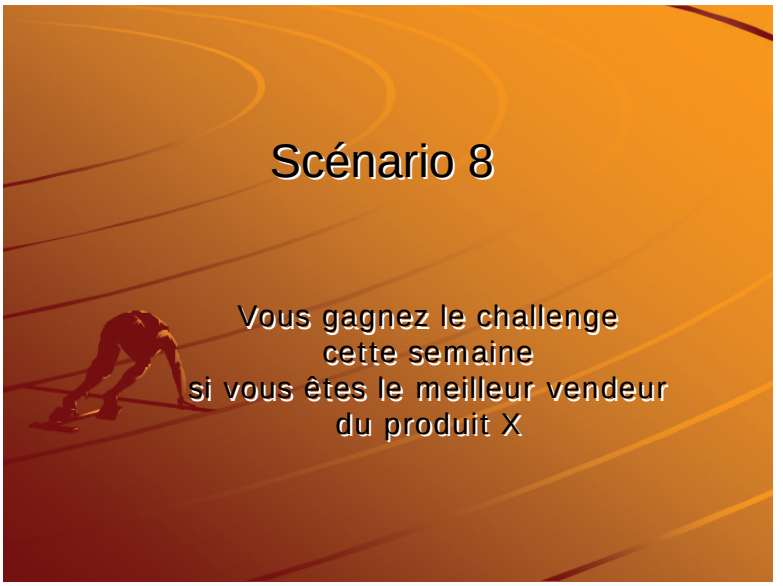
L'administration de la simulation

Compte tenu du nombre limité de variables, les données sont collectées selon la méthode des profils complets. Concernant le recueil des données, nous avons choisi un classement des profils par ordre décroissant de préférence.

L'ensemble des huit *stimuli* a été classé (par ordre décroissant de préférence) par un échantillon de 86 commerciaux lors d'une réunion de formation. Ces données ont été récupérées sur un questionnaire.

Figure 7.9

Exemple de scénario.



L'enregistrement des réponses

Une fois les scénarios classés (ou notés), les réponses doivent être saisies dans une base de données spécifique (selon le nombre de profils testés). Pour cela, allez dans le menu **Fichier > Nouveau > Créer**.

Nous avons un plan orthogonal à huit profils, il faut donc huit colonnes : de V1 à V8. Il est recommandé de créer une première colonne supplémentaire qui servira d'identifiant à chacun des questionnaires : c'est la colonne « id ». Il faut ensuite entrer le rang pour chacun des scénarios pour chacun des répondants. Nous obtenons une grille de 9 colonnes et 86 lignes (voir figure 7.10).

Sur la première ligne, figure le classement des huit challenges du premier vendeur interrogé. Nous observons qu'il a préféré le challenge n° 1, puis le 7e.

Figure 7.10

Enregistrement des données.

Fichier Edition Affichage Données Transformer Analyse Graphes Outils Fenêtre Aide

1 : id .

	id	v1	v2	v3	v4	v5	v6	v7	v8	var
1	.	1	7	6	4	8	2	3	5	
2	2	7	4	3	5	1	8	2	6	
3	3	8	5	4	2	6	7	3	1	
4	4	1	8	5	6	2	4	7	3	
5	5	4	7	8	1	6	3	5	2	
6	6	1	6	4	5	8	2	7	3	
7	7	2	8	7	3	5	4	6	1	
8	8	2	8	5	4	6	3	7	1	
9	9	1	7	4	6	5	2	3	8	
10	10	1	7	6	2	5	3	8	4	
11	11	4	8	7	2	6	3	5	1	
12	12	3	8	6	1	7	2	5	4	
13	13	4	6	7	1	8	2	3	5	
14	14	2	8	7	3	5	4	6	1	
15	15	6	1	2	7	3	5	4	8	
16	16	1	8	4	7	2	5	6	3	
17	17	1	6	7	3	4	2	8	5	
18	18	2	5	6	3	7	1	8	4	
19	19	2	8	5	4	3	6	7	1	

Vous devez sauvegarder cette base de données dans le répertoire du document contenant les données du plan orthogonal : **Fichier** > **Enregistrer sous** > **C :** …

À ce stade, nous avons le plan orthogonal généré par SPSS et les données recueillies et transcrites dans le fichier précédent. Nous utilisons une macrocommande pour réaliser l'analyse conjointe.

La commande de l'analyse conjointe

Afin d'exécuter l'analyse conjointe avec SPSS, il faut passer en mode « Syntaxe ». Pour cela, allez dans le menu **Fichier** > **Ouvrir** > **Syntaxe**.

Pour écrire la macrocommande, entrez les huit paramètres suivants :

- CONJOINT. Indique au logiciel le type d'analyse à réaliser.
- PLAN. Indique l'emplacement du document qui contient le design orthogonal.
- /DATA. Indique l'emplacement de la base de données.
- /SUBJECT. Indique le nom de la colonne qui sert à identifier les questionnaires : id.
- /RANK. Indique l'étendue des colonnes où se trouvent les valeurs accordées aux différents profils.

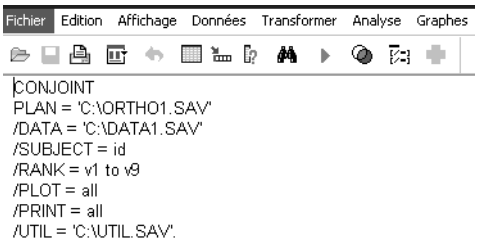
On spécifie ici le nom de la première et de la dernière colonne.

- /PLOT et /PRINT. Servent à générer les outputs.
- /UTIL. Indique l'emplacement du document contenant l'utilité de chacun des profils pour chacun des questionnaires.

Respectez bien les signes utilisés dans l'écriture de la macrocommande. L'oubli d'un seul « \ », « : » ou du point « . » à la fin de la commande empêche le logiciel de réaliser l'analyse (voir figure 7.11).

Figure 7.11

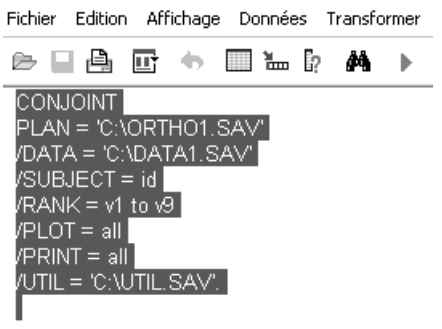
Macrocommande.



Sélectionnez ensuite l'ensemble des syntaxes et cliquez sur la flèche noire dans la barre d'outils (voir figure 7.12).

Figure 7.12

Exécution de la macrocommande.



L'analyse conjointe est réalisée. Un message apparaît qui indique que la procédure a bien fonctionné. Dans le cas contraire, vérifiez l'écriture correcte de la macrocommande de syntaxe.

Les résultats de l'analyse conjointe apparaissent dans la boîte Résultats du logiciel SPSS. Nous allons voir comment les interpréter.

L'interprétation de l'analyse conjointe

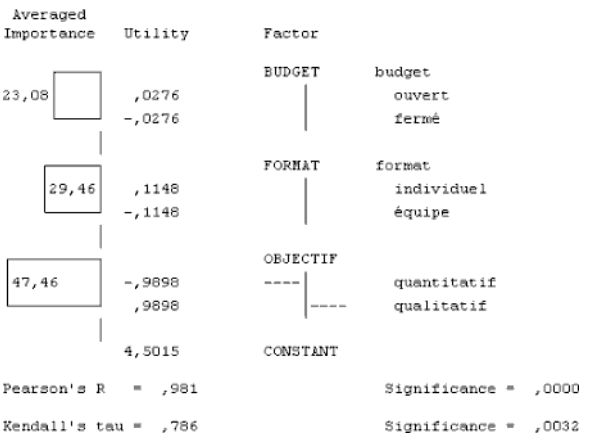
Les résultats de l'analyse des mesures conjointes se lisent individuellement, pour chaque observation, puis, à la fin de tous les résultats individuels, pour les résultats globaux.

Les utilités des modalités de facteurs sont obtenues par décomposition des scores de préférence. Pour valider les utilités ainsi obtenues, il faut comparer l'ordre dérivé des utilités de chaque combinaison de modalités avec l'ordre fourni par les préférences. Cela permet de s'assurer que les estimations des utilités fournies par les mesures conjointes permettent bien de prédire les préférences exprimées par les individus. Le coefficient tau de Kendall ou le rhô de Spearman fournissent une mesure du degré d'association ou de corrélation entre les ordres dérivés des utilités et ceux provenant des préférences. Ils varient entre 0 et 1. Plus le coefficient est proche de 1, plus les ordres associés sont proches.

Dans notre exemple (voir figure 7.13), le coefficient de concordance (tau de Kendall) qui teste l'homogénéité des préférences des individus à l'égard des challenges étant de 0,78 (proche de 1), les résultats globaux peuvent donc être acceptés.

Figure 7.13

Résultats de l'analyse des mesures conjointes.

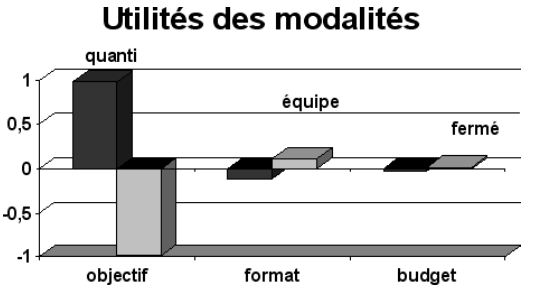


Les modalités qui ont les utilités moyennes les plus fortes sont respectivement : l'objectif quantitatif, le format en équipe et le budget fermé. La figure 7.14, qui expose les utilités moyennes et l'importance de chaque attribut, met en évidence le poids de l'objectif dans les choix du challenge, devant le format et le budget. Si aucun des trois attributs n'est négligeable, le type d'objectif est deux fois plus important que les deux autres.

Le choix des commerciaux se porte en premier lieu sur le type d'objectif du challenge, facteur très important pour eux, qui préfèrent les objectifs quantitatifs. Ensuite, apparaît le format de compétition, qu'ils apprécient en équipe. Enfin, le budget du challenge compte moins, les commerciaux inclinant pour un budget fermé avec un classement plutôt que pour un quota à atteindre.

Figure 7.14

Résultats de l'analyse des mesures conjointes (bis).



Cette analyse des mesures conjointes permet donc de souligner le type de concours idéal pour les commerciaux interviewés dans cette expérimentation : un challenge quantitatif, organisé en équipe et avec un budget fermé.

Résumé

Le comportement des individus, et en particulier des consommateurs, vis-à-vis des produits résulte d'opérations complexes qui mettent en jeu perceptions et préférences. Pour réussir la conception d'un objet, il est donc utile d'évaluer ces préférences et de modéliser les jugements des individus. L'analyse conjointe résout ce type de problème.

Cette méthode repose sur l'idée que l'importance relative accordée à un attribut par les consommateurs et les utilités qu'ils attachent aux niveaux des attributs peuvent être déterminées lorsqu'ils évaluent des profils d'objets, construits à partir de ces attributs et de leurs niveaux. Il faut donc commencer par identifier les attributs et les niveaux clés pour construire les *stimuli*. Les plans fractionnés, générés par SPSS, permettent de réduire le nombre de profils à tester. La variable dépendante est généralement une préférence estimée par un score ou un classement.

Les résultats mettent en exergue les attributs les plus importants dans l'évaluation d'un objet et les niveaux préférés à l'aide des utilités partielles. Ils permettent ainsi de dégager le produit optimal aux yeux des individus.

Pour aller plus loin

Dussaix *et al.*, *L'Analyse conjointe, la statistique et le produit idéal*, Ceresta, 1992.

Liquet J.C, *Cas d'analyse conjointe*, Broché, 2001.

Louviere J.J, *Analyzing Decision Making: Metric Conjoint Analysis*, Sage, New-York, 1998.

Exercices

Exercice 1 Étude sur la consommation de thé – création de plan d'expérience et de scénario

Énoncé

Vous êtes chef de produit chez Lipton, on vous confie le lancement d'un nouveau thé. Dans ce cadre, vous cherchez à connaître les préférences des consommateurs. Après avoir animé une réunion de groupe, vous cherchez à évaluer quatre éléments importants :

- la température (chaude, tiède, froide) ;
- le sucre (pas de sucre, un sucre, deux sucres) ;
- la force (fort, moyen, léger) ;
- l'ajout de citron (avec ou sans).

1. Créez le plan orthogonal à l'aide du logiciel SPSS.
2. Créez les *stimuli* du test de produit.

Solution

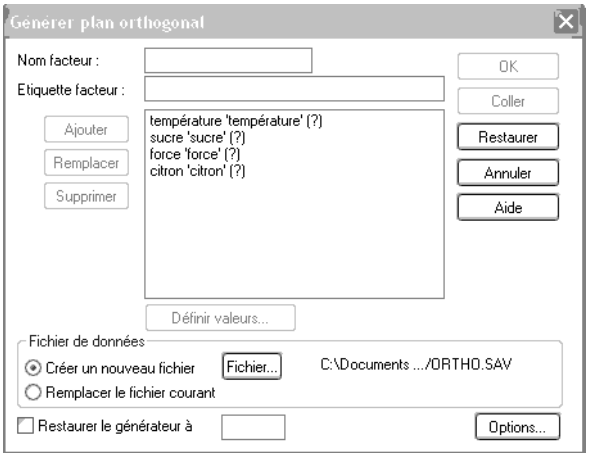
1. La création du plan orthogonal

Allez dans le menu **Données** > **Plan orthogonal** > **Générer**. Dans la boîte de dialogue qui apparaît (voir figure 7.15), tapez le nom et l'étiquette de chacun des attributs ou facteurs.

Inscrivez température dans le champ **Nom facteur** et **Etiquette facteur**. Cliquez sur le bouton **Ajouter** pour ajouter les autres facteurs toujours dans la case **Nom facteur** et **Etiquette facteur**, inscrire : sucre, **Ajoutez**, force, **Ajoutez**, citron et **Ajoutez**.

Figure 7.15

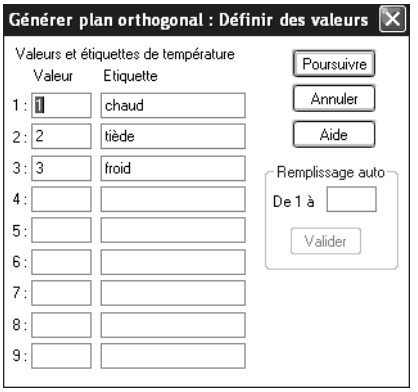
La création du plan orthogonal.



Pour définir les niveaux de chacun des facteurs, sélectionnez chaque facteur et cliquez sur le bouton **Définir valeurs**.

Figure 7.16

Enregistrement des niveaux d'attributs.



Attribuez des valeurs à tous les niveaux de chacun des attributs. Pour la température du thé, 1 pour chaud, 2 pour tiède, 3 pour froid (voir figure 7.16).

Cliquez ensuite sur le bouton **Poursuivre** pour revenir à la boîte de dialogue et refaites cette action pour chacun des attributs.

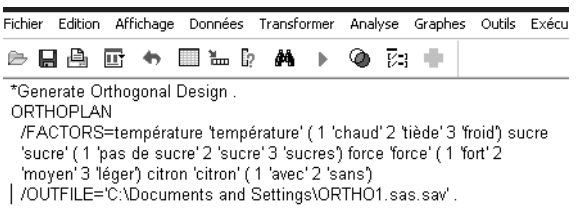
Pour le sucre, la valeur 1 pour « pas de sucre », 2 pour « un sucre », 3 pour « 2 sucres ». Pour la force du thé, la valeur 1 pour « fort », 2 pour « moyen », 3 pour « léger ». Enfin pour le citron, la valeur 1 pour « avec citron », 2 pour « sans citron ». Cliquez enfin sur **Poursuivre**.

Lorsque tous les attributs et leurs niveaux sont saisis, indiquez où vous allez placer le plan orthogonal (voir figure 7.17). Cliquez sur **Fichier** et choisissez un emplacement où le plan peut être récupéré.

Ensuite, cliquez sur le bouton **Coller** de la même boîte de dialogue (voir figure 7.18) afin d'avoir la syntaxe de la macrocommande.

Figure 7.17

Copie de la syntaxe.



Cliquez enfin sur **OK** (pour cela, revenez au menu **Données** > **Plan orthogonal** > **Générer**). Le message « Un plan est généré correctement avec 9 cartes » doit apparaître dans l'onglet **Résultats**.

Récupérez le plan orthogonal là où vous l'avez enregistré.

2. La création des stimuli pour le test de produit

Le plan orthogonal créé est un plan fractionné. En effet, nous aurions dû tester $3 \times 3 \times 3 \times 2 = 54$ profils. Or, notre plan nous permet une importante économie d'évaluations de $54 - 9 = 45$ expériences.

Le profil 1 correspond à un thé froid, comptant un sucre, léger avec du citron.

Le profil 2 correspond à un thé froid, comptant deux sucres, fort et sans citron. Etc.

Figure 7.18

Sauvegarde du fichier et demande de copie de la syntaxe.

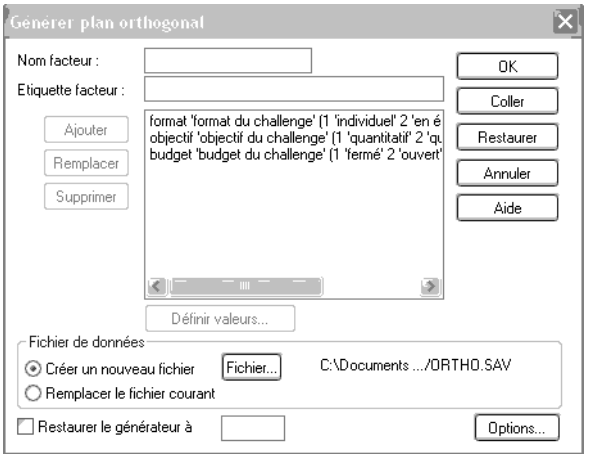
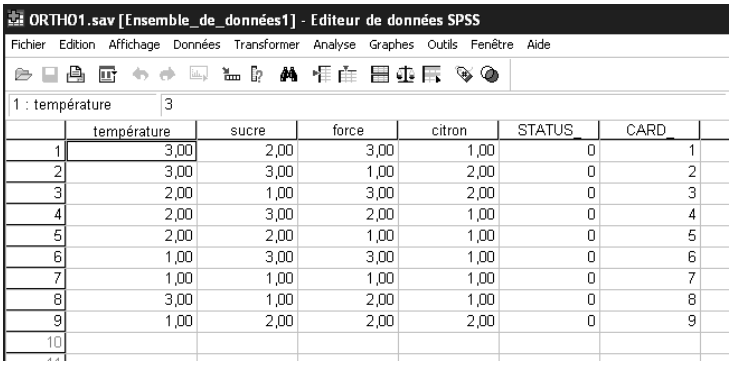


Figure 7.19

Affichage du plan orthogonal.



	température	sucre	force	citron	STATUS_	CARD_
1	3,00	2,00	3,00	1,00	0	1
2	3,00	3,00	1,00	2,00	0	2
3	2,00	1,00	3,00	2,00	0	3
4	2,00	3,00	2,00	1,00	0	4
5	2,00	2,00	1,00	1,00	0	5
6	1,00	3,00	3,00	1,00	0	6
7	1,00	1,00	1,00	1,00	0	7
8	3,00	1,00	2,00	1,00	0	8
9	1,00	2,00	2,00	2,00	0	9
10						

EXERCICE 2 ÉTUDE SUR LA CONSOMMATION DE THÉ CALCUL DES PRÉFÉRENCES DES CONSOMMATEURS

Énoncé

Des données ont été récoltées.

1. Rentrez-les sous SPSS.
2. Commandez l'analyse des mesures conjointes et interprétez les résultats de cette analyse.

Thés Individus	1	2	3	4	5	6	7	8	9
1	6	5	9	9	3	4	7	2	1
2	6	9	2	2	5	8	1	7	3
3	1	7	9	9	5	2	8	6	4
4	1	5	4	4	6	2	3	9	8
5	5	2	8	8	3	6	7	9	4

Thés Individus	1	2	3	4	5	6	7	8	9
6	2	3	4	4	1	8	7	6	9
7	9	2	3	3	7	1	5	8	4
8	6	2	1	1	5	8	9	7	3

Solution

1. L'enregistrement des données collectées

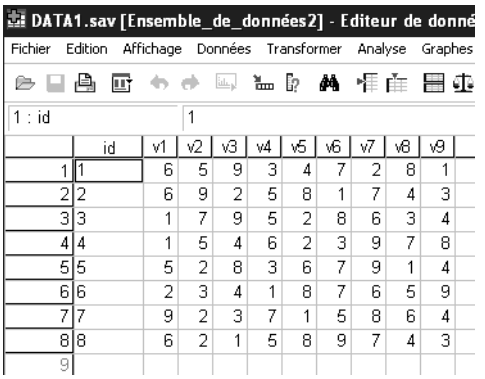
Nous avons les résultats du classement de ces neuf profils de thés. Pour saisir les données, créez un nouveau document. Allez dans **Fichier** > **Nouveau** > **Créer**.

Puisque nous avons un plan orthogonal à neuf profils, nous devons créer neuf colonnes, de V1 à V9, et une colonne d'identifiant à chacun des questionnaires : colonne « id ».

Il faut ensuite entrer le rang pour chacun des profils pour les huit répondants (voir figure 7.20).

Figure 7.20

Enregistrement des données.



	id	v1	v2	v3	v4	v5	v6	v7	v8	v9
1	1	6	5	9	3	4	7	2	8	1
2	2	6	9	2	5	8	1	7	4	3
3	3	1	7	9	5	2	8	6	3	4
4	4	1	5	4	6	2	3	9	7	8
5	5	5	2	8	3	6	7	9	1	4
6	6	2	3	4	1	8	7	6	5	9
7	7	9	2	3	7	1	5	8	6	4
8	8	6	2	1	5	8	9	7	4	3
9										

Sauvegardez la base de données dans le répertoire du document contenant les données du plan orthogonal : **Fichier** > **Enregistrer sous** > **C:** …

2. La commande de l'analyse conjointe

Afin d'exécuter l'analyse conjointe avec SPSS, allez dans le menu **Fichier** > **Ouvrir** > **Syntaxe**.

Écrivez la commande suivante (attention à l'emplacement de vos fichiers !) [voir figure 7.21].

Sélectionnez ensuite l'ensemble des syntaxes et cliquez sur la flèche noire dans la barre d'outils.

Les résultats de l'analyse conjointe apparaissent dans la partie Résultats (voir figures 7.22 et 7.23).

Le test d'homogénéité des préférences (tau de Kendall) est de 0,97, ce qui permet d'accepter les résultats globaux.

Il ressort de l'analyse conjointe que la force, la température et le sucre sont les trois attributs les plus importants dans les préférences des consommateurs de thé. Les modalités

Figure 7.21

Écriture de la macrocommande.

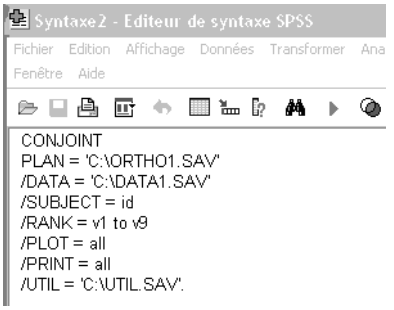


Figure 7.22

Interprétation des résultats de l'analyse conjointe.

Statistiques globales

Utilités

		Estimation d'utilité	Std. Erreur
température	chaud	-,708	,024
	tiède	,250	,024
	froid	,458	,024
sucre	pas de sucre	-,500	,024
	1 sucre	,375	,024
	2 sucres	,125	,024
force	fort	-,333	,024
	moyen	,458	,024
	léger	-,125	,024
citron	avec	-,281	,018
	sans	,281	,018
(Constante)		5,094	,018

Valeurs d'importance

température	29,807
sucre	23,500
force	30,550
citron	16,144

Score d'importance moyen

Figure 7.23

Interprétation des résultats de l'analyse conjointe (bis).

Corrélations[a]

	Valeur	Sig.
r de Pearson	1,000	,000
Tau de Kendall	,972	,000

a. Corrélations entre les préférences observées et estimées

ayant les utilités moyennes les plus fortes sont la température (les consommateurs préfèrent le thé froid), le sucre (un sucre), la force (moyen) et le citron (sans).

Chapitre 8

Communiquer les résultats

La communication des résultats constitue la dernière étape d'un projet d'étude. Ce chapitre souligne les éléments clés qui font de la communication des résultats une étape incontournable de toute analyse réussie. En effet, communiquer les résultats d'une analyse consiste à proposer au lecteur une interprétation adaptée à ses connaissances. Au-delà des éléments liés à l'interprétation des tests, que nous avons abordés tout au long des chapitres précédents, l'analyste doit être capable de formuler de manière intelligible les résultats de son étude.

1 Rédiger le rapport

La rédaction d'un rapport d'étude n'est pas directement liée à l'analyse des données. Cependant, même si le projet a été bien mené, un rapport inégal peut compromettre le succès et la valorisation d'une étude. Cette communication des résultats de l'enquête peut faire l'objet d'une ou de plusieurs mises en forme, et se trouve le plus souvent complétée d'une présentation orale que nous n'aborderons pas ici.

1.1 La structure d'un rapport d'analyse

On retrouve, en général, les mêmes éléments structurants dans un rapport, qu'il s'agisse d'un rapport d'étude ou de recherche. Ces éléments peuvent être modifiés à la marge en fonction du destinataire. Nous développerons, dans ce chapitre, la dimension professionnelle du rapport et ferons donc référence à des illustrations provenant du secteur des études de marchés.

Le rapport d'analyse fait écho au **brief de l'étude** (voir focus 1) et se structure en huit parties principales représentées à la figure 8.1.

Focus 1 — Le brief de l'étude

Le brief de l'étude (ou projet d'étude) peut servir de base à la rédaction du rapport. Il s'agit du document qui cadre l'interaction entre le client et la société d'étude, et dont la Fédération Syntec Études Marketing et Opinions a défini les grandes lignes :

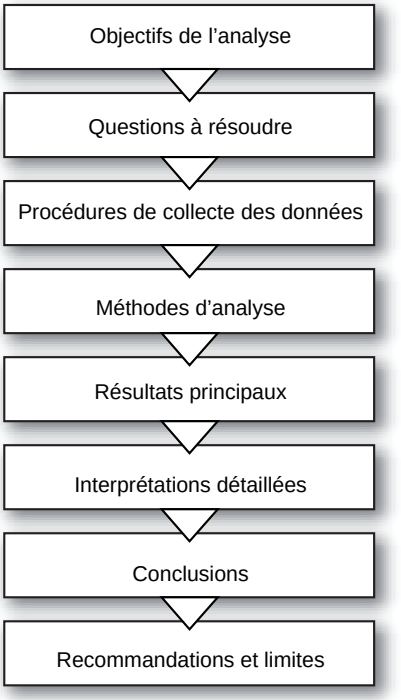
« À partir des informations fournies par le client ou à défaut d'informations complètes, en précisant ses hypothèses de travail, la société pourra remettre une proposition :

- rappelant les objectifs de l'étude et les types de résultats qui seront fournis ;
- précisant les modalités techniques :
  - mode d'échantillonnage ;
  - modalité de recueil des données ;
  - nombre et dispersion des points de sondage ;
  - plan du questionnaire ;
  - analyse et rapport ;
- indiquant les délais ;
- faisant mention du prix et des paramètres permettant de le calculer ;
- faisant état d'autres dispositions éventuelles. »

- **Objectifs de l'analyse :** cette partie est essentielle car elle résume les éléments clés de l'analyse et reprend les éléments liés à la définition du problème (tels que nous les avions abordés dans le chapitre 1). Le rappel des objectifs permet de communiquer plus précisément sur des aspects qui sont généralement maîtrisés par le chargé d'étude tels que l'approche méthodologique ou les interprétations statistiques. En les faisant apparaître clairement au destinataire, le rédacteur du rapport peut mettre en avant la résolution progressive de ces objectifs de départ. Dans le cadre de notre enquête fil rouge sur le point de vente, on pourrait formuler notre analyse en la résumant ainsi : « Qui sont les clients du futur point de vente ? ».

Figure 8.1

Structure d'un rapport d'étude.



- **Questions à résoudre :** il s'agit ici d'établir les questions principales qui permettent d'avancer dans la satisfaction aux objectifs de l'analyse. Dans notre exemple, il s'agissait, dans un premier temps, de décrire les clients en fonction des variables d'identification (âge, sexe, revenus, etc.), de les classer en grands segments homogènes, puis enfin de comprendre leurs comportements (intention d'achat par exemple).
- **Procédures de collecte des données** : cette rubrique reprend les données utilisées pour les besoins de l'étude. Il faut en préciser la source s'il s'agit de données secondaires, ainsi que les caractéristiques principales (nombre d'observations, de variables, etc.). Dans le cas de données primaires, le rédacteur du rapport devra apporter un soin particulier à la description des procédures de collecte. Quelle méthode d'échantillonnage ? Quel instrument de mesure ? Comment les données brutes ont-elles été préparées ? Dans notre exemple sur le point de vente, les données sont issues d'un questionnaire administré en face-à-face à 400 répondants, clients de l'enseigne.
- **Méthodes d'analyse** : les méthodes employées et les stratégies d'analyse sont présentées à ce stade. Dans un premier temps, les tests de nature descriptive (tris croisés, comparaisons de moyenne, etc.), puis les tests liés à la représentation des données (analyse factorielle, typologie), et enfin les tests de nature explicative (régressions, analyses de variance, etc.). Dans notre étude, nous avons tout d'abord décrit les clients par le biais des principales variables d'identification, puis mis en place deux analyses des corrélations multiples pour établir les variables liées au montant moyen dépensé et à l'intention d'achat des futurs clients du point de vente.
- **Résultats principaux** : cette partie, la plus importante du rapport, peut être constituée de plusieurs sous-parties. Pour plus de clarté, il est recommandé d'organiser ces sous-parties en fonction de thèmes énoncés dès le départ. Dans le cadre d'une étude d'image

de marque par exemple, le premier niveau de résultats consiste à évaluer les scores de notoriété obtenus, puis à les rapprocher des résultats par groupes d'attributs pour identifier les attentes des consommateurs. Les résultats principaux doivent répondre point par point aux objectifs de l'analyse et apporter une solution à la demande du client de l'étude. À la question de départ « Qui sont les clients du futur point de vente ? », nous avons répondu en plusieurs séquences : la description des variables clés, la définition des segments de clients, l'identification de déterminants de l'intention d'achat, etc.

- **Interprétations détaillées** : les interprétations détaillées ont pour objet d'apporter un éclairage technique aux résultats et d'éclairer le lecteur du rapport sur la démarche méthodologique. On trouve, dans cette rubrique, l'explication des tests mis en œuvre, les hypothèses à respecter, l'interprétation des indicateurs de significativité des tests, entre autres. Ces éléments techniques doivent être présentés de manière intelligible en fonction du profil du lecteur. En effet, ceux-ci n'ont pas tous la même expertise en termes d'échantillonnage ou de tests statistiques, et le chargé d'étude doit veiller à formuler ces éléments de manière à être compris de tous les destinataires du rapport.
- **Conclusions** : cette partie est essentielle au client de l'étude et a pour objet de lui fournir tous les éléments de synthèse utiles pour la valorisation et l'utilisation de l'étude. Une étude récente sur la situation de la grande distribution en Europe présentait ses conclusions autour de quatre grands défis à relever : les défis de la grande consommation, les défis de la gestion des ressources humaines, les défis de la relation client et les nouveaux chantiers informatiques.
- **Recommandations et limites** : les recommandations accompagnent la présentation des résultats et représentent l'engagement de l'analyste dans la réponse au problème posé par l'entreprise commanditaire. La recommandation n'est pas systématique et suppose une expertise de la part de l'analyste, ou à tout le moins que le problème posé ait été analysé suffisamment en profondeur. En principe, le fait de réaliser l'étude suppose que l'on se soit informé au préalable sur le secteur, que l'on ait été *briefé* par le client et, en conséquence, que l'on est en mesure de dépasser la simple description et de s'impliquer dans la résolution du problème managérial. Enfin, les limites ayant pesé sur l'étude doivent être mentionnées (temps, budget, moyens, etc.). Ce retour sur les conditions de réalisation de l'analyse et sur les précautions que le commanditaire devra prendre lors de l'utilisation des résultats est un élément important qui doit prévenir toute extrapolation hasardeuse, sans toutefois minimiser les résultats de l'étude.

1.2 Les points clés du rapport

La qualité d'un rapport d'analyse se juge sur deux dimensions principales : la précision et l'intelligibilité.

La précision représente la qualité du rapport en termes de crédibilité des résultats. Le document doit établir de manière précise la pertinence des choix effectués en matière de méthode, d'analyse et de résultats, ce qui présuppose que les données collectées soient de qualité, que les analyses soient valides et fiables, et que les résultats soient correctement interprétés. Ce tryptique (données, analyses, résultats) est au cœur de la démarche d'analyse et doit naturellement être envisagé de manière conjointe, ces trois dimensions établissant collectivement la crédibilité de l'étude. Plus exactement, pour que le rapport soit précis, les données doivent l'être aussi ; le rédacteur devra prendre des précautions importantes quant à la manipulation des données, la définition, l'exécution et l'interprétation

des tests statistiques. Autrement dit, aucune erreur n'est tolérable dans un rapport d'analyse. Erreurs de calcul ou de syntaxe, fautes d'orthographe, maladresses conceptuelles et terminologiques ne sont que quelques exemples de ce manque de précision qui pénalise la crédibilité d'un rapport.

La clarté de l'expression, la logique du raisonnement, la rigueur de l'expression et de l'analyse sont les fondements naturels du second critère de qualité d'un rapport : **l'intelligibilité**. Le lecteur du rapport peut éprouver des difficultés de compréhension – et par suite des doutes quant à la qualité du travail effectué – lorsque le raisonnement n'apparaît pas clairement (des hypothèses de travail insuffisamment mises en avant par exemple), ou lorsque la présentation manque de précision (une méthode d'échantillonnage qui reste confuse). Les « croyances » des commanditaires jouent également un rôle dans la bonne compréhension des résultats de l'enquête. Dans notre enquête sur le point de vente, si les responsables de l'enseigne estiment (*via* des études internes, des reportings de vente, etc.) que 50 % des clients ont l'intention d'acheter dans le nouveau magasin, et que les résultats de l'enquête sont éloignés de cette prévision, le rédacteur devra justifier cet écart. L'intelligibilité à ce niveau participe de la pédagogie : cet écart peut être dû à une mauvaise compréhension de la question ou de l'échelle de mesure, à un biais d'échantillonnage ou à une erreur dans la prise en compte des non-réponses. Le rédacteur devra répondre aux interrogations du client sur ces écarts et **expliquer ses résultats**.

2 Mettre en valeur les résultats

L'objet du rapport d'analyse est d'apporter la réponse attendue par le commanditaire de l'étude et de constituer une référence. Une fois produit, ce rapport aura une existence propre, passera de mains en mains auprès de nombreuses parties prenantes. La mise en valeur des résultats permet de prolonger cette durée de vie et autorise une utilisation globale ou partielle du rapport par les différents lecteurs. La version de SPSS dont nous nous servons (V15.0) propose quelques outils d'amélioration de la qualité des tableaux et des graphiques pour mettre en valeur les résultats de l'analyse.

2.1 Maîtriser les tableaux

Les tableaux doivent systématiquement être numérotés, comporter un titre clair mentionné dans le texte (par exemple : « le tableau 8.2 illustre les effectifs de la variable marital »), et ne doit retenir que la partie la plus significative des données. La maîtrise des tableaux est un exercice délicat, qui suppose de bien maîtriser la manière dont on souhaite communiquer les résultats, et qui repose sur la distinction vue plus haut entre résultats principaux et interprétations détaillées. Il existe, en effet, des niveaux de résultats simples facilement compréhensibles par le lecteur, et d'autres, plus complexes, qui exigent des compléments et des éclairages. La figure 8.2 illustre un résultat simple (description du statut marital des répondants de l'enquête sur le point de vente), alors que la figure 8.3 montre le tableau d'un résultat plus complexe (analyse de corrélation sur une série de variables de la même enquête).

Dans le cas d'une présentation de résultats complexes, la lecture du tableau suppose un certain nombre d'éléments facilitant l'interprétation. Les astérisques (**) et la mention en

Figure 8.2

Présentation d'un résultat simple.

Quel est votre statut marital?

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Célibataire	93	23,3	23,3	23,3
	En couple	263	65,8	65,8	89,0
	Autre (divorcé(e), veuf(ve)	44	11,0	11,0	100,0
	Total	400	100,0	100,0	

Figure 8.3

Présentation d'un résultat complexe.

Corrélations

		Montant moyen dépensé par mois	Intention d'achat dans le point de vente	Quel est votre niveau d'étude?	Taille du foyer	Revenus du foyer
Montant moyen dépensé par mois	Corrélation de Pearson	1	,333**	,666**	,038	,833**
	Sig. (bilatérale)		,000	,000	,445	,000
	N	400	400	400	400	400
Intention d'achat dans le point de vente	Corrélation de Pearson	,333**	1	,502**	,018	,283**
	Sig. (bilatérale)	,000		,000	,719	,000
	N	400	400	400	400	400
Quel est votre niveau d'étude?	Corrélation de Pearson	,666**	,502**	1	-,064	,635**
	Sig. (bilatérale)	,000	,000		,205	,000
	N	400	400	400	400	400
Taille du foyer	Corrélation de Pearson	,038	,018	-,064	1	-,002
	Sig. (bilatérale)	,445	,719	,205		,974
	N	400	400	400	400	400
Revenus du foyer	Corrélation de Pearson	,833**	,283**	,635**	-,002	1
	Sig. (bilatérale)	,000	,000	,000	,974	
	N	400	400	400	400	400

**. La corrélation est significative au niveau 0.01 (bilatéral).

bas du tableau apportent un complément important permettant de ne retenir que l'information pertinente, à savoir les variables les plus fortement corrélées entre elles (niveau d'études et intention d'achat par exemple). Il est préférable de présenter ces tableaux complexes en annexe du rapport et de communiquer directement les résultats dans une section du rapport : « le niveau d'études, le montant moyen dépensé et les revenus sont corrélés à l'intention d'achat ».

SPSS

SPSS permet de produire un grand nombre de tableaux personnalisés. La maîtrise s'acquiert en manipulant progressivement l'interface des tableaux dans l'éditeur de résultats, comme nous avons pu le voir tout au long des chapitres. Les tableaux personnalisés de SPSS s'obtiennent par la procédure suivante : **Analyse > Tableau > Tableaux personnalisés…** comme le montre la figure 8.4.

Il est toutefois utile de connaître certains éléments afin de gagner en efficacité dans ce type de tâche. Les tableaux produits dans l'éditeur de résultats de SPSS sont des **tableaux pivotants** qui autorisent une très grande flexibilité en termes de formatage et de présentation des résultats. On obtient un tableau pivotant dans SPSS en double-cliquant sur le tableau dans l'éditeur de résultats, opération faisant apparaître un menu supplémentaire (**Tableau Pivotant**) dans le menu de l'éditeur de résultats (voir figure 8.5).

Attention toutefois, car les possibilités de présentation sont directement liées au type de variable utilisée (voir à ce sujet le chapitre 2). Le générateur de tableaux se fondera donc sur l'étiquette de la variable (nominale, ordinale, échelle) que vous avez définie au préalable. Le fait de ne pouvoir générer le tableau désiré est souvent dû à une variable mal étiquetée.

La première fonction qui peut être utile dans la présentation d'un tableau est la fonction **Empiler** de SPSS. L'empilement (tout comme les autres fonctions tableaux que nous allons voir) s'obtient par le menu **Tableaux personnalisés…**, et consiste à affecter deux ou plusieurs variables en ligne ou en colonne. Vous pouvez faire glisser les variables simultanément en ligne ou en colonne, ou bien l'une après l'autre. Dans l'exemple ci-après, nous avons transformé un tableau, dans lequel la variable *Age* était en ligne et la variable *Sexe* en colonne, en un tableau où les deux variables sont en ligne, comme le montre la figure 8.6.

La fonction **Empiler** s'avère très utile lors d'enquêtes pour présenter des résultats d'échelles de mesures. Un concept comme la confiance accordée à la marque, par exemple, est

Figure 8.4

Fonction tableaux personnalisés.

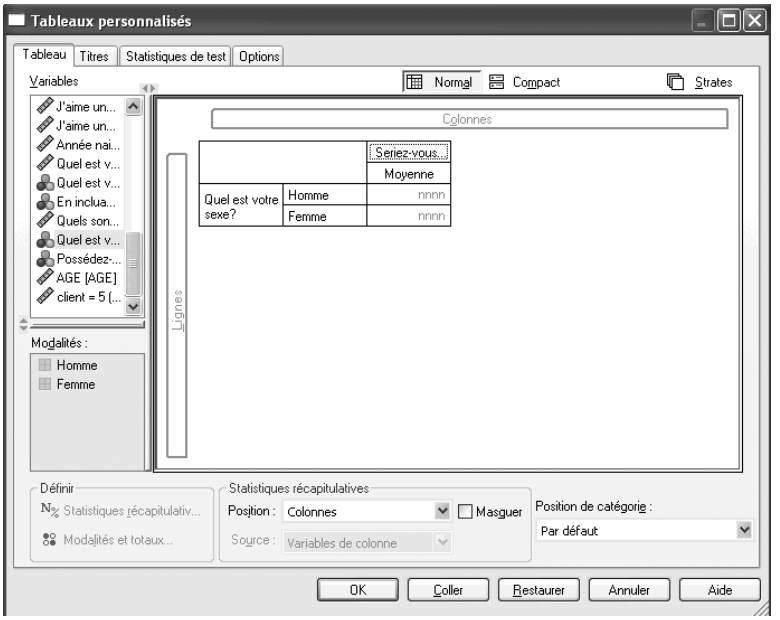


Figure 8.5

Fonction tableau pivotant.

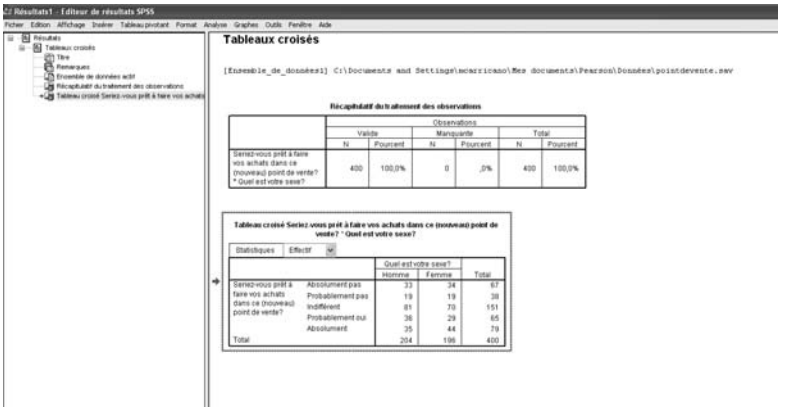


Figure 8.6

Présentation de la fonction Empiler.

		Sexe	
		Homme	Femme
Catégorie d'âge	Moins de 25	108	134
	25 à 34	276	351
	35 à 44	309	370
	45 à 54	221	260
	55 à 64	136	184
	65 et +	178	301

		Effectif
Sexe	Homme	1232
	Femme	1600
Catégorie d'âge	Moins de 25	242
	25 à 34	627
	35 à 44	679
	45 à 54	481
	55 à 64	320
	65 et +	479

mesuré par une série de variables dont on peut ainsi empiler les résultats pour en obtenir une vision exhaustive.

Plusieurs possibilités sont également disponibles à partir des tableaux croisés que nous avons abordés au chapitre 2. Pour rappel, les tableaux croisés s'obtiennent par la procédure suivante : **Analyse > Statistiques descriptives > Tableaux croisés…** La procédure de l'**Emboîtement** consiste à insérer une variable dans la même dimension d'un tableau croisé afin d'apporter un niveau de résultat supplémentaire. La figure 8.7 montre l'emboîtement de la variable *Sexe* dans la variable *Catégorie d'âge*.

Figure 8.7

Présentation de la fonction Emboîtement.

			Effectif
Catégorie d'âge	Moins de 25	Homme	108
		Femme	134
	25 à 34	Homme	276
		Femme	351
	35 à 44	Homme	309
		Femme	370
	45 à 54	Homme	221
		Femme	260
	55 à 64	Homme	136
		Femme	184
	65 et +	Homme	178
		Femme	301

À partir du menu tableaux personnalisés, vous pouvez demander des strates afin d'ajouter une dimension de profondeur à vos tableaux et créer ainsi des « cubes » tridimensionnels comme le montre la figure 8.8.

Figure 8.8

Présentation de la fonction Strates.

Sexe Femme

		Effectif
Catégorie d'âge	Moins de 25	134
	25 à 34	351
	35 à 44	370

Sexe Homme

		Effectif
Catégorie d'âge	Moins de 25	108
	25 à 34	276
	35 à 44	309
	45 à 54	221
	55 à 64	136
	65 et +	178

La fonction **Tableau** permet de contrôler les totaux et sous-totaux, les pourcentages les fréquences, afin de choisir la présentation optimale de vos résultats. Il est possible de modifier à volonté l'aspect d'un tableau en créant, par exemple, un modèle qui reprend l'ensemble des propriétés définissant l'aspect du tableau. On accède aux modèles de tableaux en double-cliquant sur le tableau dans l'éditeur de résultats et en sélectionnant dans le menu **Format > Modèles de tableaux…**

2.2 Améliorer les graphiques

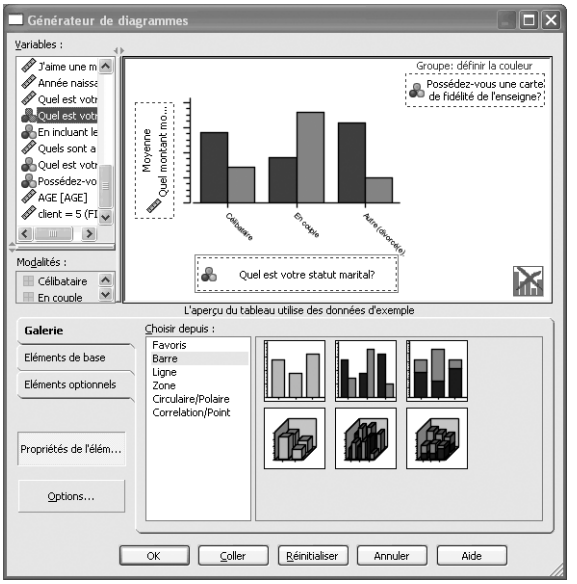
Les graphiques jouent eux aussi un rôle crucial dans la présentation des résultats. Ils enrichissent le contenu du rapport, à condition d'être présentés à bon escient. Cela implique le bon choix du modèle de graphique en fonction du test mis en œuvre et un emploi pertinent : complément d'un tableau ou d'un résultat présenté dans le texte, amélioration de la lisibilité et de la fluidité du rapport, etc. Les graphiques peuvent être considérablement enrichis grâce à SPSS. Ils sont générés de diverses manières : par le biais des principaux tests statistiques ou par l'utilitaire de diagramme que nous présentons dans cette section. Nous aborderons trois types de graphiques : les diagrammes en bâtons, les diagrammes en secteurs et les diagrammes de dispersion qui sont les plus utilisés.

SPSS

On obtient le générateur de diagramme par la procédure suivante : menu **Graphe > Générateur de diagramme…** La procédure fait apparaître la boîte de dialogue de la figure 8.9.

Figure 8.9

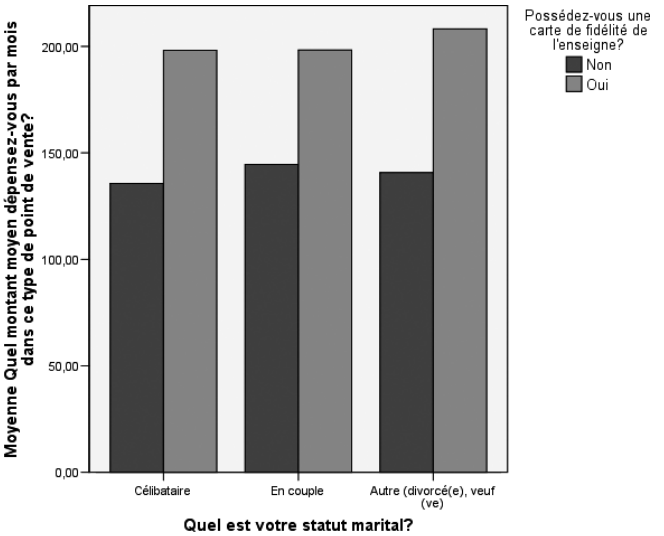
Boîte de dialogue Générateur de diagramme.



Le générateur de diagramme affiche simultanément une seconde fenêtre (**Propriétés de l'élément**) qui vous permet notamment d'afficher dans un menu déroulant des statistiques que vous pouvez insérer dans le diagramme (des effectifs par exemple). Vous n'êtes donc pas obligé de passer par un tableau pour créer un graphique de statistiques descriptives pour une ou plusieurs variables. On démarre l'utilitaire en faisant glisser l'icône représentant le diagramme envisagé dans le canevas, c'est-à-dire la large zone qui s'affiche au-dessus de la galerie (l'onglet activé sur la figure 8.9). On fait ensuite glisser les variables vers les zones de déplacement de l'axe : il existe une variable de type échelle en colonne (Quel montant moyen dépensez-vous par mois dans ce type de point de vente ?), une variable de type nominal en ligne (Quel est votre statut marital ?) et une seconde variable de type nominal (Possédez-vous une carte de fidélité de l'enseigne ?), afin de grouper les répondants dans ce diagramme en bâton groupé. Nous obtenons le diagramme de la figure 8.10, qui fait apparaître l'importance de la possession d'une carte de fidélité.

Figure 8.10

Diagramme en bâtons juxtaposés.



Vous pouvez modifier à tout moment le diagramme ou choisir un autre type de diagramme pour représenter les mêmes données. Pour modifier un élément du diagramme, activez l'éditeur de diagramme en double-cliquant dessus et utilisez le menu **Affichage > Propriétés…** Ce menu vous permettra de modifier le texte du diagramme, la couleur et le motif de remplissage des bâtons, d'ajouter du texte (par exemple un titre ou une annotation), etc.

Il existe de nombreuses autres possibilités de modification. Nous en verrons quelques-unes en présentant deux autres types de diagrammes : les graphiques en secteur et les graphiques de dispersion. On peut, par exemple, masquer certaines modalités d'un graphique en secteur. Pour créer un graphique en secteur, faites glisser un graphique dans le générateur de diagramme (dans SPSS, il se nomme diagramme circulaire/polaire). Nous allons représenter de cette manière les goûts musicaux des répondants en représentant la variable *progradio*. Après avoir fait glisser le graphique en secteur dans le générateur de diagramme, cliquez sur le menu **Options** qui vous permet d'inclure ou d'exclure des observations. On peut ainsi exclure du graphique les non-réponses au questionnaire. Pour exclure des modalités sous-représentées, ou que l'on ne souhaite pas faire apparaître, comme « ne se prononce pas » ou « ne sais pas », qui sont fréquentes dans des enquêtes en marketing, on utilise l'onglet **Modalités**. Lorsque le graphique est créé (en secteur par exemple), double-cliquez dessus pour activer l'**éditeur de diagramme**. Sélectionnez le diagramme en secteur puis dans le menu **Édition** ouvrez le sous-menu **Propriétés** comme nous l'avons vu plus haut : la boîte de dialogue de la figure 8.11 s'affiche. Activez ensuite l'onglet **Modalités** et supprimez la modalité « indifférents » qui se trouve dans la fenêtre **Exclus**.

Pour améliorer la lecture du graphique, il est intéressant d'insérer les valeurs des données. Une fois que vous avez affiché l'éditeur de diagrammes, sélectionnez le diagramme en secteur, puis dans le menu **Éléments** sélectionnez **Afficher les étiquettes de données…** Le diagramme en secteur, présenté à la figure 8.12, montre une nette préférence pour les programmes musicaux de type rock.

Il est possible de transformer les valeurs (en pourcentage par exemple) et de modifier la position de l'étiquette.

Nous allons illustrer d'autres possibilités par le biais des diagrammes de dispersion. Pour ce faire, nous changeons de jeu de données. Ouvrez le fichier « ventes_voitures.sav » que nous avons utilisé au chapitre 4. Sélectionnez le générateur de diagramme en suivant le chemin : **Graphe > Générateur de diagramme**, puis cliquez sur l'onglet **Galerie** et choisissez le diagramme de dispersion regroupée (corrélation/points) (voir figure 8.13).

Figure 8.11

Exclusion d'une modalité de variable.

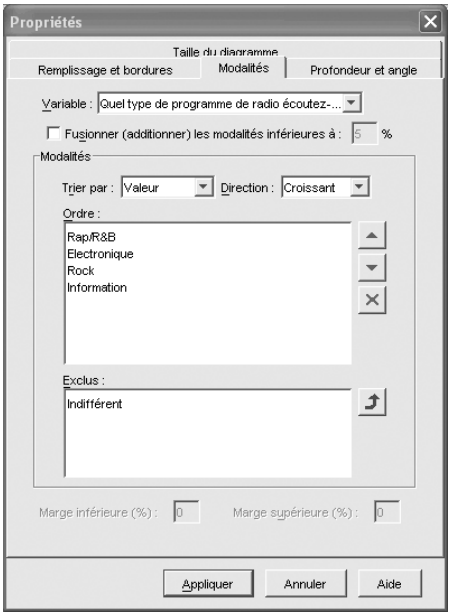
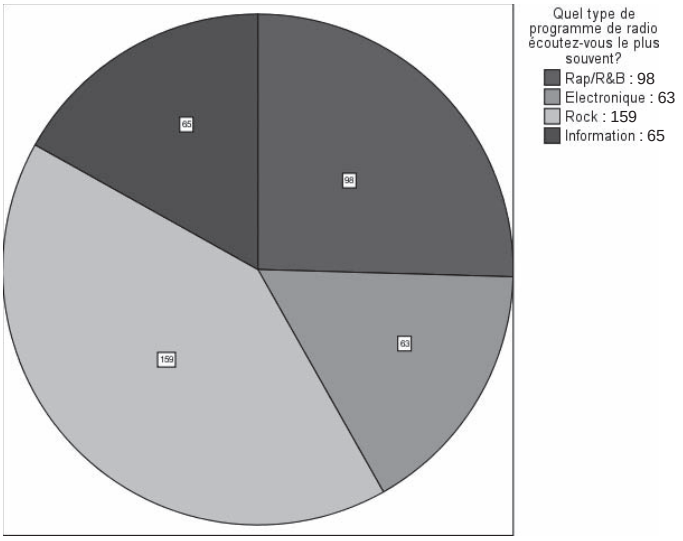


Figure 8.12

Diagramme en secteur avec valeurs.



Nous allons réaliser un diagramme de dispersion du rendement énergétique par type de véhicule avec trois variables : la consommation en colonne, le poids total à charge en ligne et le type de véhicule pour regrouper les observations. Les diagrammes de dispersion ne fonctionnent qu'avec des variables de type échelle. Nous obtenons le diagramme de la figure 8.14. qui représente la consommation du véhicule selon le poids total à charge autorisé en fonction du type de véhicule (voiture ou utilitaire).

Nous pouvons maintenant afficher une courbe d'ajustement qui permet de faire ressortir la tendance des données. La procédure est la suivante : **Éditeur de diagramme > Élements > Ajouter une courbe d'ajustement au total…** On obtient le graphique de la figure 8.15.

La courbe d'ajustement est une option intéressante car elle permet de décrire la qualité de la représentation graphique. En effet, l'option fait apparaître le coefficient de

Figure 8.13

Génération d'un diagramme de dispersion.

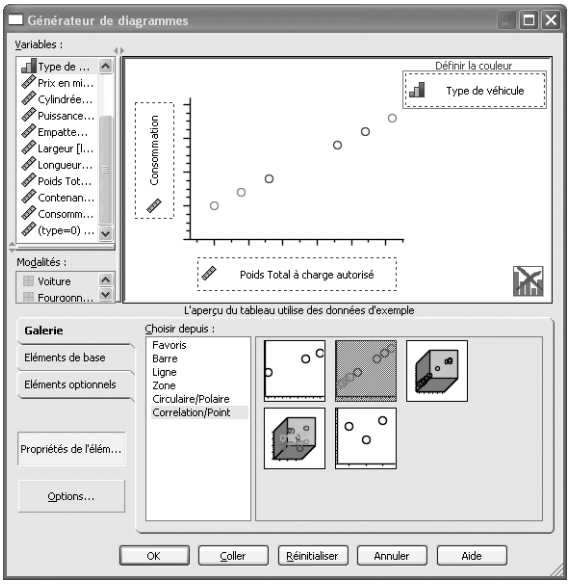
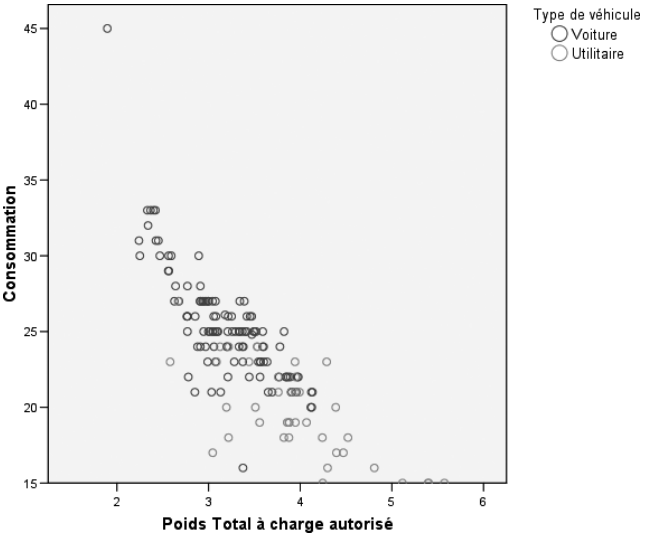


Figure 8.14

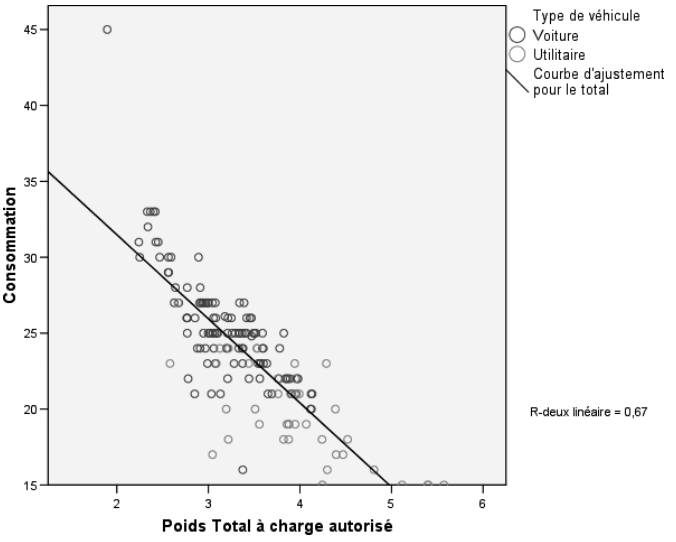
Diagramme de dispersion du rendement énergétique par type de véhicule.



détermination R^2 qui représente la proportion de variabilité de la variable dépendante (axe Y) pouvant être expliquée par la variable indépendante (axe X). Dans notre exemple, 67 % de la variabilité du rendement énergétique peuvent être expliqués par le poids du véhicule.

Figure 8.15

Diagramme de dispersion avec courbe d'ajustement.



Résumé

La préparation du rapport fait écho au brief de l'étude et reprend les grandes étapes de l'analyse des données. Son principal objectif consiste à mettre en lumière les réponses possibles au problème managérial posé. Un rapport de qualité doit être à la fois précis et intelligible, au sens où il doit présenter de façon claire pour toutes les parties prenantes de l'étude des résultats justes. La mise en valeur des résultats dans le rapport passe par la maîtrise des tableaux et des graphiques. Nous avons passé en revue dans ce chapitre les utilitaires de SPSS pour générer des tableaux pivotants et des diagrammes de qualité, mais l'utilisateur doit s'exercer, expérimenter pour pouvoir en découvrir toutes les facettes.

En guise de conclusion, et pour paraphraser Stefan Sweig : « il ne sert à rien d'éprouver les plus beaux sentiments si l'on ne parvient pas à les communiquer ».

Exercices

Exercice 1 Étude de l'impact d'une campagne de publicité

Énoncé

Afin d'observer l'impact des campagnes de publicité de ses annonceurs, un magazine de cinéma a mis en place une étude auprès d'un échantillon de 200 lecteurs représentatifs de la structure sociodémographique du lectorat. Il s'agit d'un questionnaire auto-administré, trois semaines après parution, aux seuls lecteurs ayant acheté eux-mêmes le magazine, l'ayant lu ou parcouru. Les répondants sont évalués sur des scores de reconnaissance (se souvenir avoir vu l'annonce), des notes d'agrément (de 1 à 10) et leur opinion globale (ce qui a plu ou moins plu). L'enquête permet de collecter des renseignements généraux sur l'influence du format, des emplacements, du volume publicitaire, et d'étudier plus précisément la relation entre l'impact de l'annonce et l'agrément du magazine. Les résultats doivent permettre aux marques de situer leur annonce par rapport aux standards établis pour les annonces de même format et de même secteur produit, aux annonces de la marque déjà parues les années précédentes dans le magazine, et aux annonces d'un univers de marques concurrentielles également présentes dans le support.

1. Présentez succinctement les résultats du rapport de l'étude par le biais des éléments clés que vous souhaitez faire apparaître.
2. Proposez quelques idées pour améliorer la précision et l'intelligibilité du rapport.
3. Si l'on devait présenter une extraction des résultats à un annonceur, comment devrait-on procéder ?

Solution

1. Il peut être pertinent de démarrer le rapport par un rappel de la méthodologie employée (échantillon de 200 lecteurs représentatifs, auto-administration du questionnaire dans un délai de trois semaines après la parution, etc). La rubrique suivante peut par exemple présenter des résultats globaux de type évolution des annonces, des formats, etc. Des diagrammes en bâtons ou en lignes enrichiront la présentation à ce niveau. Les résultats principaux peuvent être organisés de la manière suivante :

- **mémorisation des annonces par type de format** : on peut présenter ici les performances en termes de mémorisation par type de format par le biais d'un tableau empilé reprenant les scores par formats en pourcentage ;
- **opinion des lecteurs sur les annonces** : si des questions ouvertes ont été utilisées lors de l'enquête, on pourra utiliser ici des verbatims (des citations de lecteurs) pour mettre en lumière certains résultats quantitatifs de la première rubrique ;
- **influence du produit** : un tableau empilé ou un tableau par strates peut être utilisé pour ce type de résultats, faisant apparaître dans une colonne les effectifs d'annonces, dans une autre colonne le secteur (banque, automobile, etc.), puis le format, et enfin les taux de mémorisation ;

- **influence de l'emplacement de l'annonce** : plusieurs tableaux ou graphiques peuvent distinguer les résultats en fonction de la place de l'annonce dans le magazine (4e de couverture, premier ou deuxième cahier, etc.) ou des rubriques du magazine ;
- **influence du volume publicitaire** : le nombre d'annonces (marques et concurrents) peut être présenté ici afin de resituer les performances dans l'univers concurrentiel.

2. Pour améliorer la précision du rapport, il est important de noter certains éléments méthodologiques en fonction des résultats présentés : sur combien d'annonces ont été effectuées les scores de mémorisation, sur quelle période, pour quels produits, quelles marques en concurrence, etc. ? Les questions posées peuvent parfois être reprises, dans le texte ou en note de bas de page. En termes d'intelligibilité, il est recommandé d'utiliser la fonctionnalité Modèle de tableau ou de diagramme afin de définir un format de présentation qui vous convienne. On pourra, dans ces modèles de tableaux, faire apparaître systématiquement les résultats globaux (réponses totales par exemple) dans une couleur et les résultats les plus importants dans une autre. Lorsqu'on a recours à un tableau empilé, ce qui est fréquent pour des résultats de type descriptif à une enquête, il peut être intéressant également de traiter les variables principales avec une typologie et une couleur distinctes afin de bien identifier les différentes rubriques.

3. Le rapport peut présenter en premier lieu des résultats généraux liés à la performance de la marque : son taux de mémorisation, la satisfaction exprimée au regard de la qualité de l'annonce, etc. La présentation de visuels de la campagne de la marque serait judicieuse à ce niveau du rapport. Toujours dans l'optique de questions ouvertes, des verbatims indiqueraient l'opinion des lecteurs, ce qui leur a plu, déplu, etc. Pour présenter la relation entre satisfaction à l'égard de l'annonce et la mémorisation de l'annonce, on pourrait insérer une carte perceptuelle générée par une analyse factorielle des correspondances, comme nous l'avons vu dans le chapitre 3. Ce type de représentation est très largement utilisé dans les rapports d'analyse.

Exercice 2 Étude d'un lectorat

Énoncé

Cet exercice a pour objet de vous familiariser avec les tableaux et les diagrammes. Ouvrez le fichier « pointdevente.sav ». La maîtrise des tableaux et des diagrammes peut vous faire gagner un temps précieux lors de la rédaction du rapport.

1. Représentez les réponses à la variable *intention* qui estime l'intention d'achat dans le nouveau point de vente au moyen d'un tableau. Effectuez les choix nécessaires pour présenter le tableau le plus clair possible.
2. Pour faire écho à l'exercice 1 qui traite de l'univers de la presse, représentez sous forme de diagramme la variable *rubrikpress* qui fait ressortir les rubriques de presse quotidienne le plus souvent lues par les répondants. Vous devez produire un graphique attractif et précis.

Solution

1. Pour créer le tableau, suivre la procédure : **Analyse** > **Statistiques descriptives** > **Effectifs**… On obtient le tableau de la figure 8.16.

Figure 8.16

Effectifs de la variable *intention*.

Seriez-vous prêt à faire vos achats dans ce (nouveau) point de vente?

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Absolument pas	67	16,8	16,8	16,8
	Probablement pas	38	9,5	9,5	26,3
	Indifférent	151	37,8	37,8	64,0
	Probablement oui	65	16,3	16,3	80,3
	Absolument	79	19,8	19,8	100,0
	Total	400	100,0	100,0	

L'aspect du tableau n'est pas satisfaisant en l'état. Nous allons le modifier en passant par un modèle de tableau préexistant. Suivons la procédure indiquée dans la partie théorique du chapitre : double-cliquez sur le tableau dans l'éditeur de résultats, puis le menu **Format** > **Modèles de tableaux…** Nous sélectionnons le modèle « Avant-garde » afin de distinguer certaines rubriques du tableau qui apparaîtront en grisé. Certaines rubriques du tableau créé (pourcentage valide et pourcentage cumulé) ne sont pas directement utiles pour lire les résultats. Nous allons donc les éliminer en sélectionnant les colonnes à éliminer et en les coupant par un click droit. Enfin, lorsque votre tableau est conforme au format souhaité, vous pouvez le « copier-coller » dans votre rapport. Nous constatons le résultat à la figure 8.17.

Figure 8.17

Tableau modifié.

Seriez-vous prêt à faire vos achats dans ce (nouveau) point de vente?

		Effectifs	Pourcentage
Valide	Absolument pas	67	16,8
	Probablement pas	38	9,5
	Indifférent	151	37,8
	Probablement oui	65	16,3
	Absolument	79	19,8
	Total	400	100,0

2. Pour obtenir directement le diagramme, allez dans le menu **Graphes** > **Boîtes de dialogues héritées…** puis sélectionnez le diagramme de votre choix. La variable *rubrikpress* étant une variable nominale à 5 modalités, un graphique en secteurs est approprié. Sélectionnez **Analyse par catégories** dans la boîte de dialogue qui s'affiche (nous allons représenter les effectifs par modalité de la variable), puis **Définir**. Dans la deuxième boîte de dialogue (**Diagramme en secteurs : Groupes d'observations**), faites glisser la variable dans la rubrique **Définir les secteurs par** puis validez. Vous avez créé le graphique présenté à la figure 8.18.

Pour en améliorer l'aspect, double-cliquez sur le graphique pour ouvrir l'éditeur de diagramme, puis sélectionnez le diagramme en secteur pour l'activer. Dans le menu **Édition** > **Propriétés…** vous pouvez retravailler, par exemple, la taille du diagramme, sa profondeur et son angle, et le représenter en 3-D pour le rendre plus attractif. D'autres changements peuvent être opérés, comme nous l'avons vu, par menu **Éléments** pour afficher les étiquettes de données ou encore éclater un secteur (le plus fréquemment cité par exemple) afin de mettre en valeur les résultats. Le diagramme modifié prend l'aspect de la figure 8.19, qui est l'illustration d'une combinaison parmi d'autres. Nous vous encourageons à prolonger l'exercice sur plusieurs types de graphiques, pour bien maîtriser les tableaux et diagrammes dans SPSS : à vous de jouer maintenant !

Figure 8.18

Graphique en secteur simple.

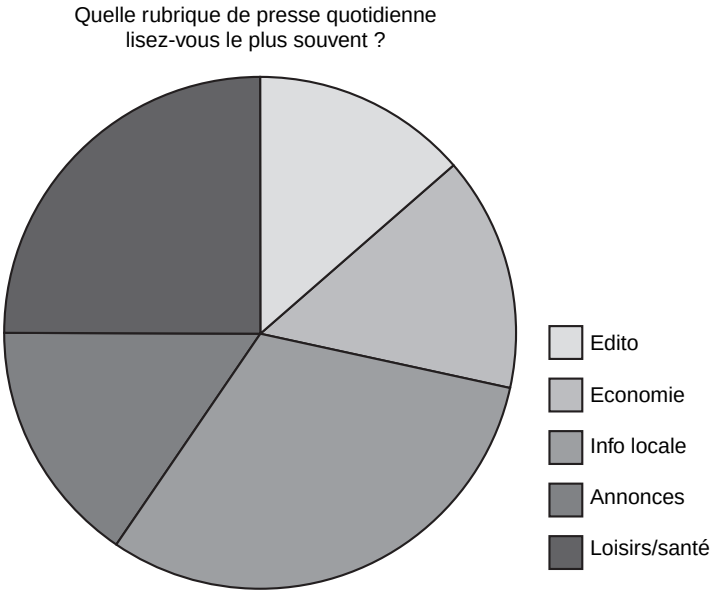
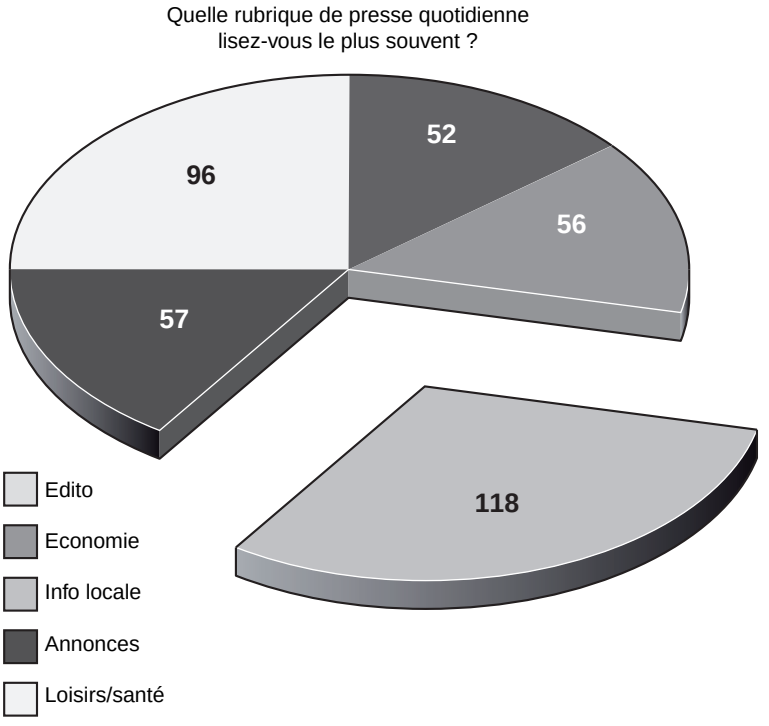


Figure 8.19

Graphique en secteur modifié.



Bibliographie générale

Churchill G. A. Jr, *Marketing Research*, 3e édition, The Dryden Press, Harcourt Brace College Publishers, 1998.

Churchill G. A., « A paradigm for developing better measures of marketing constructs », *Journal of Marketing*, 16, p. 64-73, 1979.

De Vellis R. F., *Scale developement: theory and application*, vol. 26, Sage, Thousand Oaks, 2003.

Dussaix *et al.*, *L'analyse conjointe, la statistique et le produit idéal*, Ceresta,1992.

Evrard Y., Pras B., Roux E., *Market. Études et recherches en marketing*, Nathan, Paris, 2003.

Gerbing D. W., Anderson J. C., « An updated paradigm for scale development incorporating unidimensionality and its assessment », *Journal of Marketing Research*, 25, p. 186-192, 1988.

Giannelloni J. C., Vernette E., *Les Études de marché*, Broché, Paris, 2001.

Goupy J., *Introduction aux plans d'expérience*, Dunod, Paris, 2006.

Green P. E., Srinivasan V., « Conjoint analysis in marketing: new developments with implications for research and practice », *Journal of Marketing*, 54, 4, p. 3-19, 1990.

Hair J. F. Jr., Bush R., Ortinau D., *Marketing Research*, 3e édition, Éd. McGraw-Hill-Irwin, New York, 2006.

Hair J. F. Jr., Black W. C., Babin B. J., Anderson R. E., Tatham R. L., *Multivariate Data Analysis*, 5e édition, Éd. Pearson – Prentice Hall, New Jersey, 2006.

Kozinets R.V., « The field behind the screen: Using netnography for marketing research in online communities», *Journal of Marketing*, Research, 39, 1, p. 61-72, 2002.

Lilien G. L., Rangaswamy A., De Bruyn A., *Principles of Marketing Engineering*, ISBM, State College, PA, 2007.

Liquet J. C., *Cas d'analyse conjointe*, Broché, Lavoisier, 2001.

Liquet J. C., Benavent C., *L'Analyse conjointe et ses applications en marketing*, IAE Lille, 2000.

Louviere J. J., *Analyzing Decision Making: Metric Conjoint Analysis*, Sage, New York,1988.

Malhotra N., Decaudin J. M., Bouguerra A., *Études marketing avec SPSS*, Pearson Education, Paris, 2007.

Nunnally J. C, Bernstein I. R, *Psychometric theory*, McGraw-Hill, 3e édition, 1994.

Tenenhaus M., *Méthodes statistiques en gestion*, Dunod, Paris, 2006.

Tenenhaus M., *Statistiques : méthodes pour décrire, expliquer, prévoir*, 2e édition, Dunod, Paris, 2006.

Tenenhaus M., *Méthodes statistiques en gestion*, Dunod, Paris, 2006.

Thiétard R. A., *Méthodes de recherche en management*, Dunod, Paris, 1999.

Tufféry S., *Data Mining et statistique décisionnelle*, Éd. Technip, Paris, 2007.

Vernette E., *Techniques d'étude de marché*, Vuibert, Paris, 2000.

Index

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V-W

Sciences de gestion

Synthèse de cours & exercices corrigés

Manu Carricano est enseignant-chercheur à l'INSEEC Paris où il est responsable des majeures Marketing. Il enseigne le marketing et les études de marchés en licence et master.
Fanny Poujol est maître de conférences à l'IAE de Valenciennes et chercheur associé à l'INSEEC. Elle enseigne le marketing et la méthodologie.

Analyse de données avec SPSS®

Ce livre a pour objectif d'amener à découvrir tout le potentiel de l'analyse des données à travers de nombreux exemples et exercices d'application, situés principalement dans le champ du marketing.

Progressif et pédagogique, il s'articule autour des étapes clés d'une analyse de données : la définition de la problématique, la description des données, la validation des instruments de mesure. La suite du livre met l'accent sur le choix d'une méthode d'analyse, qu'elle soit descriptive (tris croisés, analyse factorielle) ou plus technique (ANOVA, régression, analyse conjointe). Le dernier chapitre traite de la rédaction du rapport, élément essentiel de la communication des résultats.

Le livre inclut de nombreux exemples illustratifs et applications. La plupart de ces dernières font appel à SPSS afin que le lecteur se familiarise avec ce logiciel. Il pourra ainsi appliquer ses connaissances théoriques et mettre en pratique une démarche d'analyse.

Ce livre s'adresse aux étudiants de premier et de second cycle (IUT, BTS, universités et écoles de commerce). Rappel méthodologique sur la réalisation d'une analyse de données et outil concret d'utilisation de SPSS, il sera également utile aux chargés d'études en activité.

La collection Synthex propose aux gestionnaires et aux économistes de découvrir ou de réviser une discipline et de se familiariser avec ses outils au travers d'exercices résolus.

Chaque ouvrage présente une synthèse pédagogique et rigoureuse des techniques et fondements théoriques, qu'une sélection d'exercices aux corrigés détaillés permet d'assimiler progressivement. Le lecteur, étudiant ou professionnel, est ainsi conduit au cœur de la discipline considérée, et, *via* la résolution de nombreux problèmes, acquiert une compréhension rapide et un raisonnement solide.

Direction de collection :
Roland Gillet, professeur à l'université Paris 1 Panthéon-Sorbonne

Dans la même collection :

- **Analyse financière et évaluation d'entreprise,** S. Parienté
- **Performance de portefeuille,** P. Grandin *et al.*
- **Création de valeur et capital-investissement,** M. Cherif et S. Dubreuille
- **Contrôle de gestion,** Y. de Rongé et K. Cerrada
- **Économétrie,** É. Dor
- **Finance,** A. Farber *et al.*
- **Marketing, une approche quantitative,** A. Steyer *et al.*
- **Mathématiques appliquées à la gestion,** A. Szafarz *et al.*
- **Probabilités, statistique et processus stochastiques,** P. Roger
- **Stratégie,** A. Desreumaux *et al.*
- **Les enquêtes par questionnaire avec Sphinx,** S. Ganassali

ISBN : 978-2-7440-4075-7

PEARSON

Pearson Education France
47 bis, rue des Vinaigriers 75010 Paris
Tél. : 01 72 74 90 00
Fax : 01 42 05 22 17
www.pearson.fr