MARKETING ■ COMMUNICATION

Daniel CAUMONT • Silvester IVANAJ

Analyse des données

Bibliothèque en pdf

facebook

- Méthodes et outils statistiques
- Applications à la prise de décision
- Cas d'entreprise détaillés

DUNOD

MANAGEMENT SUP

Daniel Caumont
Silvester Ivanaj

Analyse des données

- Méthodes et outils statistiques
- Applications à la prise de décision
- Cas d'entreprise détaillés

DUNOD

Tout le catalogue sur
www.dunod.com



ÉDITEUR DE SAVOIRS

Conseiller éditorial : Christian Pinson

Mise en page : Belle Page

Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.
Le Code de la propriété intellectuelle du 1er juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée.
Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).



© Dunod, 2017

11, rue Paul Bert, 92240 Malakoff
www.dunod.com

ISBN 978-2-10-075851-7

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).
Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

Table des matières

Introduction 1

1 Comment construire et décrire les variables 3

Section 1 Construire le fichier des données à traiter 4

Section 2 Décrire les variables 15

2 Comment comparer des échantillons 35

Section 1 Méthodologie de la comparaison et conditions d'application 36

Section 2 Passer du résultat statistique à la décision managériale : traitement du cas « Cirage » 57

3 Comment exploiter un tableau de contingence 79

Section 1 Exploiter un tri croisé 81

Section 2 L'analyse des correspondances pour exploiter un tableau de contingence 93

4 Comment simplifier l'information avec l'analyse factorielle 115

Section 1 L'analyse factorielle des données métriques 118

Section 2 L'analyse factorielle des données non-métriques 143

5 Comment construire et valider une typologie 161

Section 1 Phase préparatoire : construire le profil de similarité 164
Section 2 Phase opératoire : obtenir différentes partitions 175
Section 3 Phase confirmatoire : valider une typologie 184

6 Comment réaliser et valider des prévisions 199

Section 1 La régression linéaire simple (RLS) 200
Section 2 La régression linéaire multiple (RLM) 221
Section 3 La prévision fondée sur des séries temporelles 241

Orientations bibliographiques 267

Tables statistiques 269

Index 275

Introduction

Avec le développement des systèmes d'information, le volume des données en gestion sollicitées dans la prise de décision n'a cessé d'augmenter. Parallèlement, la facilité d'accès aux logiciels de traitement des données est susceptible d'aider le manager à tirer profit de cette multitude pour opérer les meilleurs choix possibles. Encore faut-il maîtriser ces techniques d'analyse statistique pour associer avec pertinence l'outil de traitement des données disponibles à l'objectif managérial visé.

De ce fait, devant la profusion d'informations qu'il faut analyser avant de prendre une décision, l'étude statistique prend une importance toute particulière et doit être considérée comme l'une des matières fondamentales du cursus de l'étudiant en gestion afin qu'il soit au mieux préparé aux futurs métiers qu'il aura à exercer.

Il existe de nombreux ouvrages traitant des outils statistiques. Mais nombre d'entre eux privilégient leurs aspects conceptuels et théoriques sans toujours les resituer dans le cadre concret d'une problématique et d'une démarche liées à la prise de décision dans divers champs de l'entreprise (en marketing, ressources humaines et finance principalement). Or, le jeune décideur, autant que l'étudiant en formation, est assez souvent confronté à la difficulté du choix des outils utiles et pertinents pour l'aider dans la décision qu'il doit prendre.

C'est pourquoi cet ouvrage est focalisé sur la logique de la démarche à suivre pour parvenir à une décision managériale qui intègre l'usage de l'outil statistique dans la réflexion. Son orientation est donc à la fois méthodologique et opérationnelle : ce ne sont pas les aspects théoriques de la statistique qui sont d'abord traités, mais bien la logique intellectuelle et la finalité des outils présentés avec leurs conditions d'usage. Ainsi, les informations à traiter sont déterminées par la problématique managériale qui elle-même définit la problématique statistique, laquelle oriente le choix des outils d'analyse selon la nature des données, avant de conduire à la décision.

Les auteurs tentent de décrire avec simplicité l'essentiel des méthodes statistiques usuelles, en indiquant ce qui est indispensable pour bien interpréter le résultat des calculs nécessaires pour la prise de décision. Cependant, une rubrique « *Focus* » approfondit ponctuellement un aspect important tandis que la rubrique « *Pour aller plus loin* » suggère des pistes et des références au lecteur qui souhaite approfondir ses connaissances. Comme l'esprit et l'orientation de cet ouvrage sont largement guidés par l'expérience des auteurs, acquise tant dans la pratique en entreprise que dans leurs enseignements plus académiques, l'accent est mis sur l'interprétation des calculs, en s'appuyant sur des exemples de la vie d'un gestionnaire.

Ainsi, chacun des six chapitres est bâti autour d'un « *Cas d'illustration* » qui décrit une situation décisionnelle classique, exploité de telle manière que la rigueur de la réflexion managériale et celle du raisonnement statistique soient associées et se nourrissent mutuellement pour bien comprendre que c'est la réflexion managériale préalable qui motive le choix de l'outil statistique et non l'inverse. Cependant, la rigueur du raisonnement statistique implique un minimum de formalisation : les formules introduites sont limitées aux seules nécessaires pour comprendre la logique des outils présentés et pouvoir interpréter correctement leurs résultats.

Sur le plan méthodologique, la première étape est de transformer les informations disponibles en variables afin de les décrire avec des paramètres statistiques et des graphiques dont il faut comprendre la signification (chapitre 1). Nombre de décisions portent sur le résultat de tests (d'accueil ou de produit en marketing par exemple) réalisés sur des publics cibles : les tests statistiques d'inférence ou de comparaison peuvent guider dans les choix à faire (chapitre 2). Si le propre d'un manager est bien d'anticiper le futur, il a donc besoin d'outils pour prévoir son développement à court et moyen termes : les méthodes de régression et les techniques d'analyses des séries temporelles se basant sur une étude systématique de l'activité passée contribuent à cette réflexion (chapitre 6).

Dans de nombreux cas, l'aide à la décision repose sur l'analyse des liens qui existent entre des ensembles de variables extraites de bases de données plus ou moins hétérogènes mises à la disposition du gestionnaire, et de plus en plus volumineuses. Divers outils statistiques sont disponibles pour analyser les relations de concomitance ou de causalité qui peuvent exister entre ces variables. Mais ce champ est vaste : ne sont donc abordées que les techniques les plus usuelles qui répondent au plus grand nombre de situations habituellement rencontrées : pour exploiter un tableau croisé (chapitre 3), pour structurer les données avec les analyses factorielles métriques et non métriques (chapitre 4) et créer des groupes homogènes (chapitre 5). Une bibliographie indique les ouvrages de référence en matière de méthodologie et de théorie statistique pour le lecteur désireux d'accéder aux fondements mathématiques des modèles (non abordés dans ce livre). Enfin, un index permet de renvoyer aux concepts et notions dans les différents contextes où leur usage fait sens.

Chapitre 1

Comment construire et décrire les variables

OBJECTIFS

- Reconnaître la nature des variables et transformer un fichier contenant de l'information en une matière statistiquement exploitable avec un logiciel d'analyse des données.
- Transformer les données en variables pour organiser ces données en des unités d'information homogènes et cohérentes.
- Comprendre la signification et l'utilité des indicateurs statistiques pour les utiliser avec pertinence et savoir les interpréter.

SOMMAIRE

SECTION 1 Construire le fichier des données à traiter
SECTION 2 Décrire les variables

En gestion, nombre de décisions reposent sur l'utilisation d'informations qui proviennent de sources hétérogènes, internes ou externes à l'entreprise, dont le format est le plus souvent diversifié. Or, pour exploiter ces données de manière rationnelle, après s'être assuré de leur validité, il faut pouvoir les harmoniser en les restructurant pour qu'elles passent du statut d'information (centré sur leur contenu) à celui de variable (centré sur leur nature métrique ou non), d'autant plus qu'un même contenu peut parfois générer plusieurs variables.

Une fois transformées en variables, les informations brutes deviennent exploitables avec les méthodes d'analyse statistique qu'impose la problématique managériale à résoudre (tels que les outils pour la description, la comparaison, l'extrapolation, la prévision, voire la construction de modèles de décision, etc.).

Mais l'analyse statistique nécessaire au traitement d'une problématique décisionnelle n'est envisageable qu'au terme d'un processus comportant plusieurs phases, depuis la sélection des informations jusqu'à la validation des variables construites :

1. *La construction du tableau de données*
   - Sélection des informations pertinentes
   - Définition et construction des variables
2. *La vérification de la qualité du tableau de données*
   - Exactitude des informations
   - Traitement des données manquantes
3. *La description des variables avec leurs statistiques*
   - Apport des représentations graphiques
   - Usage des statistiques descriptives
   - Étude des liens entre les variables

Ce chapitre comporte deux sections. La première est consacrée à la construction et la validation du tableau de données, correspondant aux deux premières étapes ci-dessus. La seconde section présente les paramètres statistiques permettant de décrire et de vérifier tant la pertinence que la qualité du fichier constitué, avant d'effectuer tous les calculs qui répondent plus spécifiquement aux problématiques managériales à traiter et qui relèvent de diverses problématiques statistiques, soit comparative (les tests d'inférence), soit descriptive (les analyses factorielles), soit explicative (les modèles régressifs), voire prévisionnelle (l'analyse des séries chronologiques) dont traitent les différents chapitres de cet ouvrage.

Section 1 CONSTRUIRE LE FICHIER DES DONNÉES À TRAITER

Pour rester dans le cadre d'une approche pédagogique simple, ce chapitre est illustré avec un exemple d'informations primaires issues d'un questionnaire (lui-même simplifié). L'étude dont il est extrait avait pour objectif de décider de l'opportunité ou non de

proposer un nouveau service ferroviaire saisonnier pour desservir une zone touristique. À partir de cet exemple, nous tenterons d'expliciter le vocabulaire utilisé en analyse des données, avant d'exposer les étapes menant des informations initiales (le contenu du questionnaire) au tableau des données nécessaires pour utiliser les logiciels de traitement statistique.

Cas d'illustration

L'intérêt touristique du Massif Vosgien en hiver

Le Massif Vosgien, qui sépare la Lorraine de l'Alsace, et célébré par le slogan « Je vois la vie en Vosges », offre un certain nombre d'activités liées à la nature, même en hiver. En effet, si un enneigement maximal n'est pas garanti sur toute cette période, la pratique du ski nordique et alpin y reste possible, ainsi que celle de nombreuses autres activités de plein air praticables en famille. Afin d'adapter ses services aux besoins générés par ces activités touristiques sur le versant lorrain du Massif Vosgien en période hivernale, la Direction régionale de la SNCF a financé une étude ayant deux objectifs principaux :

- évaluer le potentiel touristique hivernal des vallées vosgiennes généré par la clientèle «hors régionale», c'est-à-dire extérieure aux départements limitrophes du massif ;
- disposer d'informations étayées émanant des organismes publics ou professionnels et des vacanciers eux-mêmes, afin de développer une argumentation convaincante auprès des instances de décision locales et régionales pour les sensibiliser et les impliquer dans le financement de nouveaux services de transport à développer pendant la saison hivernale. Plus particulièrement, la Direction régionale s'interroge sur l'opportunité de développer une relation directe entre Paris et le Massif Vosgien en fin de semaine pendant la saison hivernale.

Cette étude visait à recueillir nombre d'informations diverses concernant chacune des stations hivernales : (a) la diversité des activités offertes et le niveau d'animation ; (b) les capacités d'accueil en matière d'hébergement ; (c) la fréquentation actuelle et son évolution au cours des six années précédentes ; (e) les caractéristiques et le profil « vacancier » des différentes clientèles ; (f) leur appréciation des services et des activités offerts en matière d'accueil, d'hébergement, de loisirs, ainsi que sur les moyens d'accès et les possibilités de transport sur place.

Pour connaître cette clientèle « hors-région » et son niveau de satisfaction, un questionnaire détaillé fut administré au bord des pistes sur les trois zones géographiques qui couvrent l'essentiel des stations fréquentées, sur la base d'un échantillonnage aléatoire combinant les sites, les lieux, les périodes et les horaires d'enquête ; au total, 321 questionnaires ont été exploitables.

FRÉQUENTATION TOURISTIQUE DU MASSIF VOSGIEN (extraits du questionnaire originel)

Q.1 **Quelle est la durée de votre séjour dans les Vosges (en nombre de jours) ?**

_____

☞

© Dunod – Toute reproduction non autorisée est un délit.

Q.2 **Pour ces vacances, êtes-vous seul, avec un groupe organisé, ou en famille ?**

– seul 1
– avec un groupe organisé par un organisme 2
– avec un groupe composé d'amis 3
– en famille et avec des amis 4 => demander «taille du groupe»
– en famille exclusivement 5 => demander «taille du groupe»

(À poser seulement si réponse 4 ou 5 en Q.2) :
Combien de personnes composent le groupe ?

_____

Q.3 **Les vacanciers peuvent venir dans les Vosges pour pratiquer différentes activités, dont le ski. Vous, personnellement, pratiquez-vous :**

– le ski fond : oui 1 non 2
– le ski alpin : oui 1 non 2

Q.4 **Pour chacun des aspects que je vais vous énumérer, pouvez-vous me dire si vous êtes ou non satisfait en utilisant l'échelle de réponse suivante.** Vous indiquez : 0 (zéro) si vous n'êtes «pas du tout satisfait» et 5 si vous êtes «tout à fait satisfait» ?

– l'ensemble des activités offertes par la station : 0 1 2 3 4 5
– la location de matériel (ski, raquettes ou autre) : 0 1 2 3 4 5
– l'accueil qui vous a été fait : 0 1 2 3 4 5
– la qualité d'enneigement des pistes : 0 1 2 3 4 5

Q.5 **Pensez-vous revenir dans les Vosges pour d'autres vacances ?**

certainement pas	probablement pas	pourquoi pas ?	très probablement	très certainement
1	2	3	4	5

Q.6 **Je vais vous citer des opinions concernant le Massif Vosgien. Vous nous indiquerez dans quelle mesure vous êtes plutôt en accord ou en désaccord avec chacune d'elle**

		Pas du tout d'accord	Pas tellement d'accord	Ni d'accord ni pas d'accord	Assez d'accord	Tout à fait d'accord
a	Le plaisir des Vosges, c'est que neige ou pas neige, on ne s'ennuie jamais car il y a une grande diversité d'activités possibles.	1	2	3	4	5
b	Le massif des Vosges est vraiment adapté pour des vacances en famille ou entre amis.	1	2	3	4	5
c	On ne vient pas dans les Vosges d'abord pour le ski, mais pour l'ambiance, l'accueil, la diversité et l'accessibilité des sites.	1	2	3	4	5

Q.7 a) - Sexe : Masculin 1 Féminin 2

b) - Âge : Quelle est votre année de naissance ? 19 ____

1 La représentativité des informations à traiter

En statistique, l'ensemble des éléments sur lequel porte l'analyse est appelé *population* (ou population statistique). Pour désigner un élément appartenant à cette population, on emploie plusieurs termes considérés comme étant synonymes : *unité statistique*, *observation* ou *individu* quand l'élément est une personne. Le nombre total des éléments composant une population statistique est appelé *taille de la population*, notée *N*, de grandeur variable.

> **Exemples de populations :** les salariés d'une firme, les entreprises du CAC40, les habitants d'une ville, les élèves d'une promotion ; l'ensemble des pièces produites par une machine sur une période donnée, etc. La taille d'une population peut donc être très variée.

Il peut être utile d'étudier de manière exhaustive tous les éléments constitutifs d'une population : c'est le principe d'un *recensement*. Cela s'impose chaque fois que la population de référence ne comporte que quelques unités statistiques ou si sa taille est petite : le fait d'échantillonner conduirait à un échantillon de taille proche de celle de la population (compte tenu d'impératifs de précision fixés). Il s'impose encore si, vu les objectifs de l'étude, le coût de mise en œuvre d'un échantillon présentant toutes les garanties de validité est pratiquement aussi élevé que celui d'un recensement. Comme les données issues d'un recensement correspondent à la population réelle, l'usage d'outils statistiques reposant sur une logique probabiliste n'est donc pas utile.

En fait, pour diverses raisons (intégralité de la population inconnue, coût de réalisation très élevé, etc.), il n'est pas envisagé d'étudier la population entière : on n'en exploite qu'une petite partie, appelée *échantillon*.

Définition

> **Échantillonnage** : procédure qui permet de sélectionner, dans une population de référence, un sous-ensemble d'éléments – des individus, des objets, voire des situations – qui vont composer l'échantillon, lequel est censé représenter cette population. La taille de l'échantillon est notée *n* et le ratio *n/N* – appelé *taux de sondage* – traduit l'échelle de l'échantillon (sans incidence majeure sur la représentativité).

Remarque

> La taille d'un échantillon est généralement inférieure à 10 000 unités et, le plus souvent, comprise entre 1 000 et 2 000 unités (nombre d'études statistiques de qualité portent sur des échantillons de taille bien inférieure, limitée à quelques centaines d'observations).

Un échantillon doit donc être *représentatif* de la population qu'il représente (appelée *population de référence* ou *population-mère*). Rappelons ici que c'est la technique d'échantillonnage qui détermine la représentativité, et non pas le taux de sondage.

© Dunod – Toute reproduction non autorisée est un délit.

À considérer les deux groupes de techniques d'échantillonnage (probabilistes et empiriques) qui existent, seules les techniques *probabilistes* permettent d'obtenir *par construction* un échantillon représentatif d'une population dont la structure est inconnue. Elles seules justifient l'usage des méthodes statistiques d'extrapolation (déterminer les caractéristiques de la population à partir des observations faites sur l'échantillon). Si l'échantillonnage par quota permet de construire empiriquement un échantillon « représentatif », c'est-à-dire à l'image de la population (en se basant sur sa structure connue), il n'est cependant pas de nature aléatoire et n'autorise donc pas l'extrapolation statistique. Le tirage aléatoire simple qui consiste à tirer de manière aléatoire n éléments parmi N dans la population-mère, est la méthode la plus efficace pour obtenir un échantillon représentatif d'une population donnée. Nombre d'outils statistiques imposent que l'échantillon étudié soit statistiquement de nature aléatoire.

2 La définition d'une variable

Les termes « information », « donnée » et « modalité », « variable » ou « caractère » sont souvent utilisés. Sont-ils équivalents ou désignent-ils des entités différentes ? Le questionnaire sur la fréquentation touristique du Massif Vosgien a pour objectif d'apporter un certain type d'information (car ce ne fut pas la seule source utilisée) sur le potentiel touristique des stations étudiées à travers l'analyse de caractéristiques factuelles (des comportements) et subjectives (des perceptions et des évaluations) recueillies auprès de vacanciers questionnés. Ces caractéristiques ou « *caractères* », telles que les pratiques sportives, le niveau de satisfaction, l'âge, le genre, etc., sont « mesurées » pour chaque répondant ; et chacune des 321 personnes de l'échantillon constitue une *observation* (*i.e.* un élément, une unité statistique).

Chacune des caractéristiques est l'objet d'une mesure qui correspond très concrètement à l'une des modalités de réponse proposée dans le questionnaire utilisé. Ces *modalités*, variant d'une question à l'autre, peuvent être différentes d'une caractéristique à l'autre.

> **Trois exemples.** *La question 2* comprend 5 modalités de réponse et chacune informe sur la manière qu'a le répondant d'organiser ses vacances : elle mesure donc un état de l'observation : par exemple, en « groupe organisé » (modalité 1) ou « exclusivement en famille » (modalité 5). La *question 4* mesure la satisfaction du répondant sur quatre aspects liés à l'usage de services accessibles dans leur station de résidence. Cette seule question concerne en fait quatre caractéristiques différentes et demande donc quatre réponses pour chaque observation, soit quatre « mesures ». Mais ces caractéristiques sont évaluées avec un même outil qui mesure l'*intensité* de la satisfaction au moyen de six modalités de réponses (de 0 à 5) ordonnées. Quant à la *question 1*, elle mesure une grandeur, celle de la durée du séjour qui s'exprime directement en nombre de jours.

Ces exemples illustrent ce qu'est une *variable* avec la diversité de ses formes : l'expression des différents états ou valeurs qu'une caractéristique (ou un caractère) peut prendre dans la population étudiée.

Définitions

Variable : Une variable statistique est une information décrivant une observation. Elle correspond à l'ensemble des modalités ou valeurs qu'une caractéristique quelconque peut prendre et qui sont susceptibles de varier d'une observation à l'autre. En analyse des données, il est d'usage de désigner une variable par une lettre majuscule, souvent X ou Y. La valeur de cette variable X mesurée sur une observation particulière est désignée par la même lettre en minuscule x_i où l'indice i indique qu'il s'agit de la $i^{ème}$ observation.

Modalité : Une modalité correspond à une valeur possible parmi l'ensemble des valeurs qui définissent une variable – les modalités de réponses dans une question par exemple ; chaque nombre utilisé symbolise, et permet d'identifier sans erreur, la valeur que prend la caractéristique pour une observation donnée. L'ensemble des modalités d'une variable contient donc toutes les valeurs que cette variable peut prendre.

Une variable résulte de la décision d'attribuer une valeur numérique à chacune des modalités qui définissent un état spécifique de la caractéristique étudiée. Ou dit autrement : une caractéristique dont un nombre définit chacune des modalités qui la composent est une variable, laquelle *mesure* un état (le sexe par exemple), une intensité (le degré d'accord avec une opinion) ou une grandeur (le nombre de jours de villégiature). Aussi, une *donnée* est le résultat d'une mesure effectuée sur une observation. L'ensemble des données est donc constitué de l'ensemble des mesures obtenues pour toutes les observations sur toutes les variables. Un *tableau de données* récapitule l'ensemble des informations qui ont été « mesurées » sur la totalité de l'échantillon étudié.

3 Les différents types de données et les échelles de mesure

C'est précisément parce que la mesure d'une caractéristique varie d'une observation à l'autre que cette entité porte le nom de variable. Dans les exemples cités, la durée du séjour, l'âge ou le sexe, l'intensité d'une opinion, etc. ne sont pas des mesures de nature comparable. D'une part, elles n'expriment pas le même type de contenu : elles sont donc de nature différente ; d'où l'existence de plusieurs catégories de données. D'autre part, on ne peut « mesurer » la variable « sexe » de la même manière que l'intensité d'une opinion. Il existe des « échelles de mesure » qui rendent compte de cette diversité et spécifient la signification qu'il faut donner à chaque mesure effectuée. Ainsi, les variables sont classées en deux catégories distinctes : les variables qualitatives et les variables quantitatives.

© Dunod – Toute reproduction non autorisée est un délit.

3.1 Les variables qualitatives ou non-métriques

Dans une variable qualitative, les nombres utilisés ne correspondent pas à une mesure vraie (i.e. une grandeur) mais représentent ou symbolisent une situation, un attribut ou un état constatés. D'où leur nom de *variable non-métrique*. Les nombres utilisés se prêtent plus aux opérations logiques que mathématiques (on ne peut calculer une moyenne par exemple).

Remarque

Dans le langage de la statistique, on utilise le terme « modalité » pour désigner les différentes valeurs qui peuvent être prises par une variable non-métrique. Par exemple, les modalités de la variable Sexe sont « Masculin » et « Féminin ». Les modalités d'une variable qualitative sont choisies pour décrire de manière exhaustive les différents états de la caractéristique représentée, entre lesquels les choix sont mutuellement exclusifs.

Il existe deux types de variables non métriques : les variables nominales et les variables ordinales.

– Dans une **variable nominale**, les nombres utilisés servent à identifier les différentes modalités de réponse retenues dont l'ordre de présentation ne répond à aucune structure particulière ; de ce fait, modifier l'ordre des modalités ne change en rien la signification de l'information.

Exemple : La variable Sexe est nominale : elle est constituée de deux modalités dont l'affectation des nombres – 1 pour « masculin » et 2 pour « féminin » – est arbitraire, mais finalement respectée par les analystes ; l'éventuelle inversion des modalités ne change pas la nature de l'information. Il en est de même pour les dix classes professionnelles définies par l'INSEE que tout analyste respecte car elles définissent les chapitres du document officiel qui détaille leur contenu.

– Dans une **variable ordinale**, les nombres utilisés définissent une hiérarchie entre les différentes modalités : la structure des réponses étant ordonnée *a priori*, ces nombres traduisent l'ordre des modalités, mais sans préciser la distance qui les sépare. De ce fait, l'ordre des modalités ne peut être interverti sans modifier, voire dénaturer, le sens de l'information recueillie.
– Par ailleurs, la valeur zéro ne peut pas exister (sauf à indiquer parfois l'absence de réponse). La valeur 1 est la plus importante : elle signifie le premier ; en conséquence, plus la valeur faciale des nombres est grande, moins la modalité correspondante est d'importance, d'où la séquence : $1 > 2 > 3$, etc.

Exemples : On peut utiliser une échelle verbale pour qualifier un individu selon son âge en définissant plusieurs niveaux, tels que : «1-bébé/2-enfant/3-adolescent/4-adulte/5-senior», mais les écarts entre les échelons sont inégaux. Et supposons qu'un analyste décide que la variable Sexe est de nature ordinale : cela signifie que 1 est supérieur à 2, et donc que 1-« masculin » est « avant » ou « plus important » que 2-« féminin »…

3.2 Les variables quantitatives ou métriques

Dans une variable quantitative, les nombres correspondent à une mesure vraie et définissent une grandeur ou une quantité : d'où leur nom de *valeurs*. Ces variables se prêtent aux opérations mathématiques usuelles. Il en existe deux types : les variables d'intervalles et les variables de proportion ou « échelle de rapport ».

– Dans une **variable d'intervalles**, les nombres servent à identifier les divers échelons – les intervalles – entre lesquels les distances sont rigoureusement égales, mais dont la valeur zéro, conventionnelle, ne constitue qu'un point de repère et non l'absence du phénomène étudié. Ce type de variable supporte donc l'usage de valeurs négatives.

> **Deux exemples.** L'échelle de température de Celsius est un exemple caractéristique d'échelle d'intervalles : le zéro est un point de repère conventionnel pour la mesure de la température (fondée sur un critère physique : le changement d'état eau/glace à 0°), mais non l'absence de chaleur ; de ce fait, on ne peut donc affirmer que 20°C est deux fois plus chaud que 10°C.
>
> Dans le cas « Massif Vosgien », l'échelle de satisfaction de la Question 4 est du même type : variant entre deux pôles opposés de 0 à 5 – « pas du tout satisfait »/« tout à fait satisfait » – ces valeurs définissent cinq intervalles supposés chacun correspondre à un même changement d'intensité de satisfaction. Par ailleurs, « 0 » exprime bien un degré de satisfaction – une totale insatisfaction en fait – et non l'absence d'évaluation de la satisfaction par le répondant.

– Dans une **variable de proportion** (ou « échelle de rapport »), le zéro est un zéro absolu qui correspond à l'absence du phénomène étudié et, de ce fait, il marque l'origine des valeurs mesurées : il n'y a pas de valeurs négatives, car la valeur 0 marque l'origine de l'échelle. Comme dans l'échelle d'intervalles, la distance entre les nombres utilisés est égale.

> **Exemple.** Dans l'échelle de température de Kelvin, le zéro marque l'absence de chaleur, physiquement définissable par l'absence de mouvement moléculaire. Dans cette logique, on peut affirmer que 400°K correspond à deux fois plus de chaleur que 200°K, soit deux fois plus chaud.

Une variable de proportion peut être continue ou discrète. *Continue*, elle peut prendre toutes les valeurs possibles comprises dans son intervalle de variation (la distance kilométrique parcourue annuellement par un véhicule) ; *discrète* (ou discontinue), elle ne peut prendre que quelques valeurs entières ou spécifiques dans ce même intervalle.

> **Exemple.** La variable mesurant le nombre de vacanciers qui composent le groupe est discrète car ses valeurs ne peuvent être qu'un nombre entier (dont 0 est exclu : un groupe est composé de plusieurs personnes). La variable « année de naissance » est aussi de nature discrète, mais on peut en déduire la variable « âge » qui est continue puisqu'elle peut s'exprimer en nombre d'années, de mois, de jours…

© Dunod – Toute reproduction non autorisée est un délit.

Les variables continues se prêtent mal à certaines analyses : comme elle peut prendre n'importe quelle valeur, le nombre de modalités peut être très élevé (tel que l'âge). Aussi, pour simplifier les calculs liés à une variable continue, on peut regrouper ces valeurs en classes. Par exemple, on va créer des classes d'âge, des classes de revenu, à la place des valeurs brutes directement relevées.

4 Construire le fichier des variables

Une fois recueillie l'information brute, il faut construire un fichier de données qui soit exploitable avec un logiciel d'analyse statistique[1]. Ce fichier organise en variables l'ensemble de ces informations. Si ces dernières proviennent d'un questionnaire (exemple du cas « Massif Vosgien »), le fichier doit contenir toutes les réponses obtenues auprès de chaque répondant. Selon son contenu, une même question peut donner lieu à la création d'une seule ou de plusieurs variables, chacune ayant alors son libellé propre.

> Dans le cas « Massif Vosgien », les sept questions qui correspondant à sept classes d'information différentes donnent lieu à la création de 15 variables (*cf.* tableau 1.1).

Le fichier des données est un tableau rectangulaire croisant lignes et colonnes. Chaque *ligne* correspond à une observation (un individu dans un questionnaire) ; une observation est caractérisée par l'ensemble des valeurs observées sur chaque variable (disposée en colonne) : il y a donc autant de lignes que d'observations (soit 321 lignes dans l'exemple traité). Par ailleurs, chaque *colonne* correspond à une variable qui contient l'ensemble des valeurs prises et réparties entre toutes les observations (il y a autant de colonnes que de variables créées). Le premier travail consiste donc à créer les variables.

4.1. Transformer l'information brute en variable

Le tableau 1.1 contient quinze variables pour un questionnaire comportant sept questions. Si, très souvent, information brute (ici, les questions) et variable se confondent, il est d'autres cas où l'information recherchée repose sur plusieurs caractères (ou *items*, à l'exemple de la Question 4) : chacun d'eux génère une variable unique ayant sa métrique spécifique. Cela conduit donc à créer autant de variables différentes qu'il y a de caractéristiques distinctes.

> **Exemple du cas « Massif Vosgien ».** L'information contenue dans les questions 1 et 5 (durée du séjour et intention de revenir) est unique. Celle contenue dans toutes les autres questions repose sur plusieurs caractères donnant lieu chacun à une variable : ainsi les Q.2, Q.3 et Q.7 génèrent chacune deux variables, trois pour la Q.6 et quatre pour la Q.4.

1. Trois logiciels seront utilisés dans cet ouvrage selon les thèmes abordés : Statgraphics, très implanté dans les entreprises et qui propose une aide à l'interprétation des résultats ; SPAD, très ergonomique et particulièrement efficace pour le traitement de données non-métriques et la classification ; SPSS, utilisé en entreprise et surtout dans le monde académique et universitaire.

© Dunod – Toute reproduction non autorisée est un délit.

Tableau 1.1 – Extrait du fichier des données du cas « Massif Vosgien »

QUESTIONS	Q.1	Q.2	Q.2	Q.3	Q.3	Q.4	Q.4	Q.4	Q.4	Q.5	Q.6	Q.6	Q.6	Q.7	Q.7
VARIABLES	var1	var2	var3	var4	var5	var6	var7	var8	var9	var10	var11	var12	var13	var14	var15
LIBELLE	DurSej	Type Groupe	Taille Groupe	SkiFond	SkiAlpin	Satis Activites	Satis LocMat	Satis Acceuil	Satis Neige	Revenir Vosges	Plaisir Vosges	Adapte Vacances	Ambiance	Sexe	Age
observation															
1	5	4	7	2	1	0	0	3	3	4	4	4	4	2	19
3	5	4	6	1	2	6	7	2	4	4	3	3	3	1	20
4	5	5	5	2	2	5	2	5	4	2	5	1	2	1	20
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
16	3	1		1	1	4	3	4	5	4	4	4	3	2	25
19	5	5	7	2	1	3	1	3	4	4	4	4	4	2	26
21	5	4	9	1	1	3	5	4	3	1	5	5	5	1	27
23	5	5	3	1	1	6	6	3	3	3	4	4	3	1	27
25	5	4	7	2	2	0	0	0	4	3	2	5	1	2	28
28	5	4	7	1	1	5	6	5	2	3	2	4	5	2	29
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
89	5	5	5	1	2	5	3	3	3	4	4	4	4	2	38
90	4	4	6	1	1	2	4	3	3	5	2	4	5	0	38
91	5	4	8	1	2	1	6	3	3	1	5	5	5	1	39
92	5	5	4	2	1	2	2	2	3	2	5	1	2	1	39
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Dans d'autres contextes, il peut être nécessaire de créer une variable unique à partir de plusieurs informations différentes.

> **Exemple.** Dans une étude publicitaire, une variable « *Crédibilité* » synthétise la crédibilité de chaque annonce telle que perçue par l'ensemble des répondants ; par construction, ses valeurs devaient varier entre 0 et 1 (0 : difficile… et 1 : facile… à croire) ; de forme $[(N+n_2)/(N+n_1)] - [n_0/N]$, où : N est le nombre total de répondants par annonce, n_0 : le nombre de répondants considérant l'annonce comme non crédible, n_1 : le nombre de position neutre et n_2 : le nombre de répondants considérant l'annonce comme crédible.

Afin de les identifier sans ambiguïté, chaque variable est alors dénommée avec un libellé propre (une étiquette) reporté dans le fichier de données. À l'exemple de l'écran d'affichage des variables sous SPSS (figure 1.1), l'usage d'un logiciel exige cette identification (colonne « nom ») ainsi que la nature de la métrique de chaque variable (colonne « mesure », ce qui implique de bien maîtriser leur métrique).

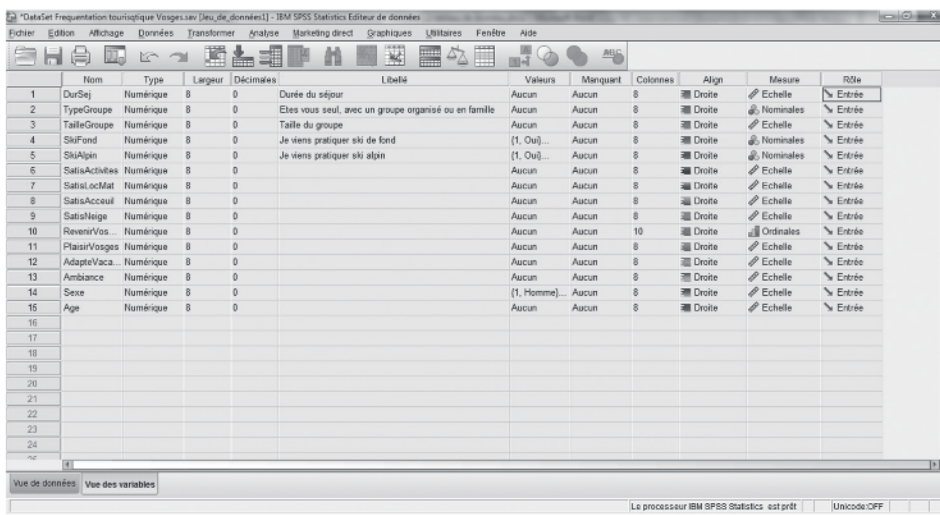


Figure 1.1 – Exemple de tableau des variables (avec SPSS)

4.2. Traiter les données manquantes

Il n'est pas rare qu'un fichier de données soit incomplet : certaines variables ne sont pas informées pour quelques observations. Si ces lacunes ne concernent pas plus de 10 % des observations et sont réparties de manière aléatoire, on peut s'en accommoder, à la condition que cela ne pose pas un problème de représentativité (indispensable pour extrapoler) ou de précision du résultat (taille de l'échantillon trop faible). Cependant, si ces données sont issues d'un questionnaire et concernent de manière systématique un petit nombre de répondants, il faut alors se demander si cela n'exprime pas à une stratégie de réponse voulue (abstention sur un thème par exemple) et s'interroger sur la possible suppression des observations incomplètes.

Remarque

Ne pas confondre les informations manquantes avec l'absence de données dans une variable due à un filtrage imposé par une variable antérieure, à l'exemple de la variable 3 dans le tableau 1.1 (taille du groupe) où l'absence de valeurs est justifiée pour toutes les observations qui, dans la variable 2, contiennent l'une des modalités comprises de 1 à 3.

Si une observation est incomplète, il existe une alternative à sa suppression : combler l'information manquante si cela ne concerne que peu de variables (et peu d'observations dans l'échantillon). C'est la solution à retenir à chaque fois que la nature des données le rend possible et si cette opération ne remet pas en cause la cohérence des informations. Deux procédures sont utilisables : soit remplacer la valeur manquante au moyen d'un calcul, soit lui substituer une donnée moyenne.

– *Remplacer la valeur manquante au moyen d'un calcul.* Cette procédure est adaptée pour traiter les données manquantes dans une série informative, à l'exemple d'une suite d'items associés à la mesure d'une attitude. Il s'agit de déterminer une valeur plausible liée à un profil de réponses existant sur d'autres items ou questions. Les valeurs observées sur l'ensemble des variables concernées sont utilisées pour déterminer quelle serait la valeur manquante la plus réaliste. Après un calcul impliquant l'usage d'outils statistiques (fondés sur la régression pour les variables métriques ou la théorie de l'information pour les variables non métriques), la valeur obtenue remplace la valeur manquante[1].
– *Remplacer la valeur manquante par une donnée moyenne.* C'est la procédure la plus simple : la valeur manquante est remplacée par la valeur moyenne de la variable observée sur l'échantillon d'observations (la moyenne si la variable est métrique ou la valeur modale si elle n'est pas métrique). L'effet de cette solution peu coûteuse est neutre puisque l'observation manquante est assimilée à une « réponse moyenne ».

Quelle que soit la solution retenue, aucune n'est exempte de biais et disposer d'un fichier de données complet d'origine reste l'idéal, d'où l'intérêt, voire la nécessité de mettre les moyens nécessaires pour obtenir un fichier de qualité.

Section 2 DÉCRIRE LES VARIABLES

L'analyse d'une variable commence par son tri à plat qui est en fait le tableau de la distribution de ses données triées selon ses différentes valeurs : cela consiste tout simplement à dénombrer les résultats obtenus. Cette analyse repose sur des statistiques qui sont des indicateurs numériques renseignant sur les caractéristiques fon-

1. Pour un traitement complet des non-réponses, consulter : Dussaix A-M. et Grosbras, J-M. : *Les sondages : principes et méthodes* P.U.F., chapitre IX ; Lebart, L. (édit.) : *La qualité de l'information dans les enquêtes*, ASU – Dunod.

© Dunod – Toute reproduction non autorisée est un délit.

damentales de la variable : sa tendance globale, la dispersion des observations, sa forme (en référence à une loi de distribution). Des graphiques permettent de visualiser ces statistiques et donnent une première impression sur la variable.

1 Construire le tableau de distribution

Définition

Tableau de distribution : tableau qui « distribue » toutes les observations entre chacune des valeurs ou des modalités prises par une variable donnée et en donne l'effectif. Il y a un tableau de distribution par variable et, donc, autant de tableaux de distribution que de variables contenues dans le tableau de données. Ce tableau est aussi nommé tri à plat.

Dans une variable non métrique, le nombre de *catégories* ou de *modalités* de la variable est le plus souvent limité à quelques unités : à chaque modalité correspond une valeur x de la variable.

Exemple. La variable 2 (Type de groupe), qui est non-métrique, comprend cinq modalités : la valeur x d'une observation donnée peut prendre l'une de ces cinq valeurs (entre 1 et 5). Le tableau de distribution de la variable distribue donc les 321 observations entre ces cinq valeurs (exprimées en effectifs et en fréquences) :

Tableau 1.2 – Tableau des effectifs pour v2-TypeGroupe (sous Statgraphics)

Modalités	Classe : valeur x	Effectif : n	Fréquence : f	Effectif cumulé	Fréquence cumulée
1- seul	1	64	0,1994	64	0,1994
2- groupe organisé	2	32	0,0997	96	0,2991
3- groupe d'amis	3	64	0,1994	160	0,4984
4- famille et amis	4	80	0,2492	240	0,7477
5- famille exclusivité	5	81	0,2523	321	1,0000

Dans une variable métrique continue, le nombre de valeurs x peut être élevé, voire trop élevé à considérer les effectifs observés. Aussi, pour construire un tableau de distribution qui ait quelque intérêt, il faut *discrétiser* la variable, c'est-à-dire la transformer pour réduire le nombre de valeurs initiales en les regroupant en « classes ». Chaque classe est affectée d'une nouvelle valeur x qui la caractérise.

Exemple. La variable « v15-âge » comporte 53 valeurs x différentes dont la fréquence oscille entre 0,31 % (soit 9 valeurs différentes de x) et 5,61 % des observations (soit 18 éléments pour une seule valeur de x). Il convient de réduire l'étendue de cette variable à quelques classes fixées en relation avec les objectifs de l'étude (classes d'âge générationnelles). Mais on peut aussi décider d'un autre nombre de classes, lié au respect d'une loi de distribution statistique (telle que la loi normale de Laplace-Gauss). Le tableau 1.3 distribue les 321 observations dans 9 classes, telles que définies selon le critère de Sturges (*cf.* rubrique « Pour aller plus loin »).

Tableau 1.3 – Tableau de distribution des effectifs pour la variable « Âge » réduite à 9 classes (v15 bis – AgeClasses)

Classes d'âge	Valeur x_i	Effectif n_i	Fréquence $f_i = \frac{n_i}{n}$	Effectif cumulé	Fréquence cumulée	Densité $d_i = \frac{n_i}{a_i}$
1 : [19 ; 25[	1	13	0,0405	13	0,0405	2,128
2 : [25 ; 31[	2	21	0,0654	34	0,1059	3,437
3 : [31 ; 37[	3	37	0,1153	71	0,2212	6,056
4 : [37 ; 43[	4	67	0,2087	138	0,4299	10,966
5 : [43 ; 49[	5	80	0,2492	218	0,6791	13,093
6 : [49 ; 55[	6	54	0,1682	272	0,8474	8,838
7 : [55 ; 61[	7	29	0,0903	301	0,9377	4,746
8 : [61 ; 67[	8	11	0,0343	312	0,9720	1,800
9 : [67 ; 75[	9	9	0,0280	321	1,0000	1,473
Total		321	1,0000			52,537
Formulation du total :		$\sum_{i=1}^{n} n_i$	$\sum_{i=1}^{n} f_i$			

Pour aller plus loin – Comment créer des classes ou discrétiser une variable continue ?

Pour réduire l'étendue d'une variable métrique continue, ou pour simplifier les calculs, les valeurs de la variable peuvent être regroupées en classes.

Les classes sont caractérisées par leur nombre (k), leur amplitude (a_i), leur centre (a_i) et leur densité (d_i). On peut construire des classes d'amplitude égale pour garder le caractère métrique de la variable. Si l'amplitude est inégale, la variable change de statut et devient non-métrique ordinale.

Transformer une variable continue en plusieurs classes (principe de la « discrétisation ») s'effectue en quatre étapes : (1) définir le nombre de classes ; (2) calculer l'amplitude des classes ; (3) définir les bornes de classe ; (4) calculer le centre des classes.

1. Déterminer le nombre k de classes :

Le regroupement des données en classes simplifie les calculs, mais entraîne une perte d'information par rapport aux données initiales (moins de détails). Il n'existe pas de loi pour déterminer le nombre optimal de classes (cette détermination est subjective), mais l'objectif est d'obtenir le meilleur compromis entre simplification des calculs et perte d'information (souhaitée aussi réduite que possible). Cependant, il existe des critères pour optimiser le nombre de classes (k) en fonction de l'effectif total (n) :

Le critère de Sturge :	Le critère de Yule :	Critère empirique :
$k = 1 + \frac{10}{3}\log(n)$	$k = \frac{5}{2}\sqrt[4]{n} = 2,5n^{0,25}$	$k = \sqrt{n}$

☞

© Dunod – Toute reproduction non autorisée est un délit.

À ces méthodes s'ajoute une pratique basée sur l'expérience qui repose sur le nombre d'observations :

Nombre d'observations (Taille n de l'échantillon)	Nombre k de classes
Moins de 50	5 – 7
50 – 100	7 – 8
101 – 500	8 – 10
501 – 1 000	10 – 11
1 001 – 5 000	11 – 14
Plus de 5 000	14 – 20

Mais le nombre de classes doit être un nombre entier : il faut donc arrondir la valeur calculée à l'entier supérieur le plus proche. Une lecture de l'histogramme des valeurs permet aussi de modifier le nombre des classes afin d'obtenir une meilleure lecture des données.

2. Calculer l'amplitude (a) des classes :

Si les classes à construire doivent être d'amplitude identique, cette amplitude (ou étendue élémentaire des classes, ou largeur des intervalles) est alors donnée par la formule : $a = \omega/k$ où : $\omega = x_{(i\ max)} - x_{(i\ min)}$ (l'étendue des données) et k = nombre de classes.

Par exemple, l'étendue des données pour la variable « âge » du tableau 1.3 vaut : ω = 74-19=55 ; d'où a = 55/9 = 6,11 ; l'amplitude interclasse est fixée à 6 (d'où une 9e classe dont l'amplitude est 8 et non 6)

3. Définir les bornes $(x_{(min)} - x_{(max)})$ de classe :

C'est déterminer pour chaque classe ses valeurs limites inférieure et supérieure. Par convention, un intervalle de classe est fermé à gauche et ouvert à droite, d'où un type d'écriture particulier : par exemple l'intervalle d'âge [49;55[se lit « valeurs de 49 ans inclus à moins de 55 ans ».

4. Calculer le centre de chaque classe :

La variable étant transformée, pour calculer les différents paramètres statistiques décrivant une variable métrique (moyenne, écart-type), il faut substituer aux valeurs initiales de la classe une valeur les représentant : c'est le milieu de l'intervalle de la classe. Par exemple, pour l'âge, l'amplitude ω étant de 6, il y a donc 5 intervalles : pour la classe 6 : [49-54[, la valeur centrale de la classe vaut donc (49+54)/2 = 51,5.

Une dernière information mesure la concentration des observations : c'est la densité (d_i) d'une classe (k_i). Elle est donnée par le rapport $d_i = n_i/a_i$ où : n_i est l'effectif de la classe i, et a_i : l'amplitude de la classe. Pour la même classe 6 : [49-54[, elle vaut 54/6,11, soit 8,838 (*cf.* tableau 1.3).

Une fois les valeurs initiales de la variable regroupées en classes, le tableau de distribution statistique en donne la nouvelle répartition. Ce tableau indique, pour chaque

☞

modalité x ou classe de données, le nombre d'éléments ou d'observations caractérisés par cette classe, comme l'illustre le tableau 1.3 : la variable « âge » étudiée peut prendre 9 valeurs pour chacune desquelles ont été calculés l'effectif et la fréquence des observations, avec leur cumul. Par exemple, la valeur 3 définit la classe qui regroupe tous les âges compris de 31 ans à moins de 37 ans, soit 37 observations et représente 11,53 % de l'effectif total.

2 La représentation graphique : une manière de visualiser les variables

Il existe différentes sortes de diagrammes et de courbes qui donnent une vision immédiate de la répartition des données entre les classes du tableau de distribution. L'œil perçoit de manière synthétique les traits dominants d'une distribution et ses détails les plus significatifs. Tableurs (type Excel) et bien des logiciels proposent une diversité de graphiques plus ou moins complexes qui permettent aussi de présenter plusieurs variables simultanément sur un même diagramme (« toile d'araignée » ou « radar » par exemple).

Ces graphiques prennent différents noms selon la nature des variables. Les plus usuels reposent sur un système d'axes (en abscisse et ordonnée) à l'exemple de ceux présentés dans le tableau 1.4 et la figure 1.2. Quel que soit le système de représentation utilisé, une légende doit spécifier sans ambiguïté le contenu et la nature des données en respectant les conventions établies : par exemple, l'abscisse porte les valeurs de la variable et l'ordonnée l'effectif ou la fréquence.

Tableau 1.4 - Les différents diagrammes de visualisation

Types de diagrammes et nom générique :	Variables non métriques ou métriques discrétisées	Variables métriques continues
Distribution des observations par modalités ou valeurs : *diagramme différentiel*	- Diagramme en bâtons - Diagramme en secteurs	- Histogramme - Graphique de densité
Cumul des observations d'une modalité (ou valeur) à l'autre : *diagramme cumulatif*	- Courbe cumulative en paliers	- Courbe cumulative linéaire

Il convient de respecter certaines règles d'usage :

– si les variables sont ***non-métriques*** : on utilise les diagrammes en bâton (« *bar-chart* ») verticaux ou horizontaux et les diagrammes en secteurs (en français « *camembert* », en anglais « *pie-chart* » : à chacun ses références culinaires !)
– si les variables sont ***métriques*** : on utilise des histogrammes et des courbes de fréquence (lorsqu'il convient de lisser les modalités sur la base des centres de classes pour représenter la forme d'une distribution continue).

© Dunod – Toute reproduction non autorisée est un délit.

2.1 Les diagrammes descriptifs

Le diagramme en bâton (pour les variables non métriques) ou l'histogramme (pour les variables métriques)

Ces graphiques dessinent la répartition des observations entre les différentes modalités de la variable étudiée. L'effectif est représenté par un trait ou un rectangle dont la hauteur ou la *surface* lui est proportionnelle. Si diagramme ou histogramme renseigne sur la forme de la distribution, il met aussi en évidence les modalités les plus prégnantes et celles de moindre fréquence, voire les modalités vides.

Ce type de représentation graphique est adapté pour comparer les différences de fréquence entre des groupes séparés. De plus, dans le cas de variables métriques, dès lors que la continuité de la mesure a été rompue (discrétisée) pour constituer des classes, l'histogramme renseigne sur le type de loi de distribution statistique approchée (loi normale ou loi de Poisson par exemple).

Le diagramme en secteurs

Plus qu'un diagramme ou un histogramme, le diagramme en secteurs permet de visualiser les rapports de proportion existant entre toutes les modalités : chaque secteur est défini par un *angle* proportionnel aux effectifs de la modalité présentée.

D'une manière générale, diagramme en bâton et histogramme fournissent différents types d'information utiles pour une première approche des données. Ils informent :

- *sur le centre approximatif des données :* sont-elles focalisées autour de la valeur centrale de la variable ou décentrées vers les classes de valeurs supérieures ou inférieures (figure 1.2). Cela donne une vision globale de la tendance des observations.
- *sur la répartition des données* entre les différentes valeurs de la variable : sont-elles fortement concentrées autour de la valeur centrale (signe d'un consensus dans la mesure d'une opinion par exemple) ou sont-elles plutôt dispersées entre toutes les valeurs (forte différenciation des observations). Observe-t-on des regroupements d'éléments autour de quelques valeurs, signalant l'existence de segments bien séparés ou, au contraire, une large dispersion des observations entre toutes les valeurs ?
- *sur la forme de la distribution* et, dans le cas d'une variable métrique, sur le degré de symétrie de la distribution des données, pouvant suggérer la référence à une loi statistique usuelle, telle que la loi normale (courbe en cloche).

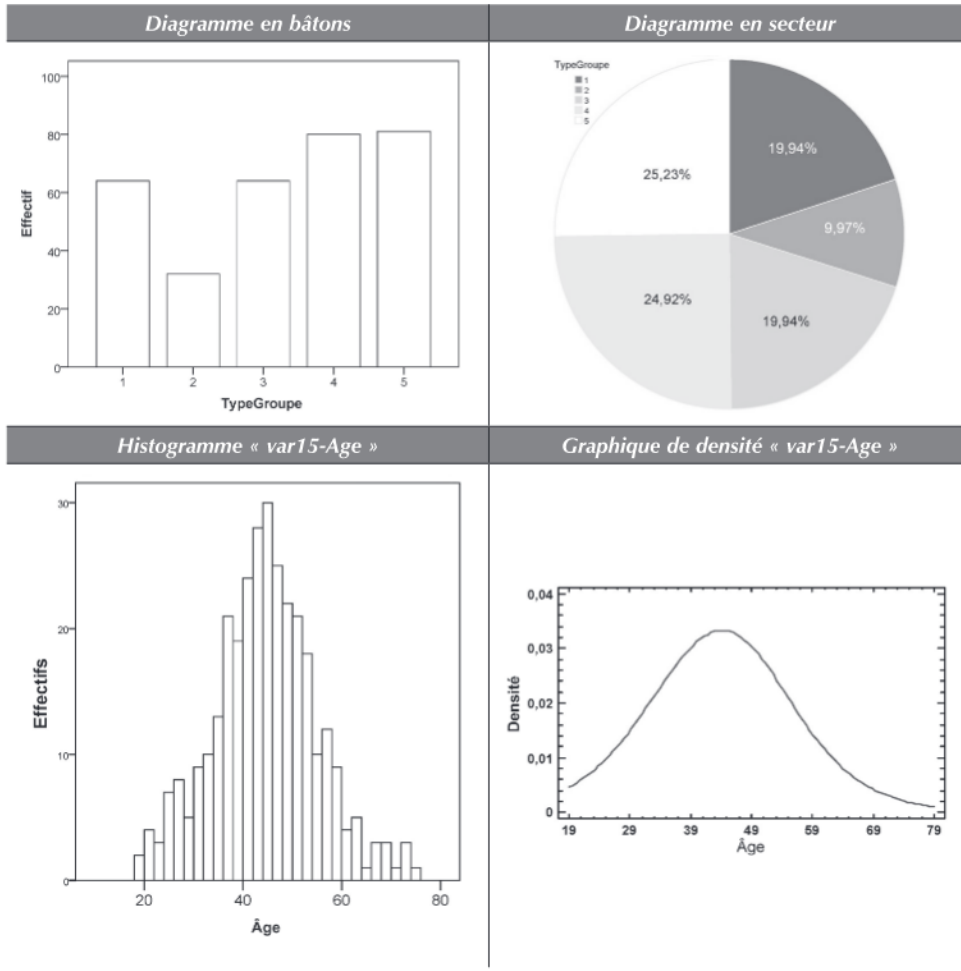


Figure 1.2 – Différents types de diagrammes (exemple « Massif Vosgien »)

2.2 Les graphiques spécifiques aux variables métriques

D'autres graphiques sont utilisés pour étudier plus spécifiquement les variables métriques, soit pour représenter la dispersion de leurs données (« boîte à moustaches » et diagramme en « tiges et feuilles ») décrits ci-dessous ; soit pour figurer des liens pouvant exister entre elles (diagrammes de dispersion et en ligne).

La boîte à moustaches (« box and whisker plot »)

C'est un diagramme de dispersion (tableau 1.5) qui renseigne sur la structure de dispersion des données d'une variable et qui permet de repérer l'existence d'observations marginales sur lesquels il convient de s'interroger. Ce diagramme est composé de deux parties : la boîte et les moustaches. La boîte est le rectangle central

© Dunod – Toute reproduction non autorisée est un délit.

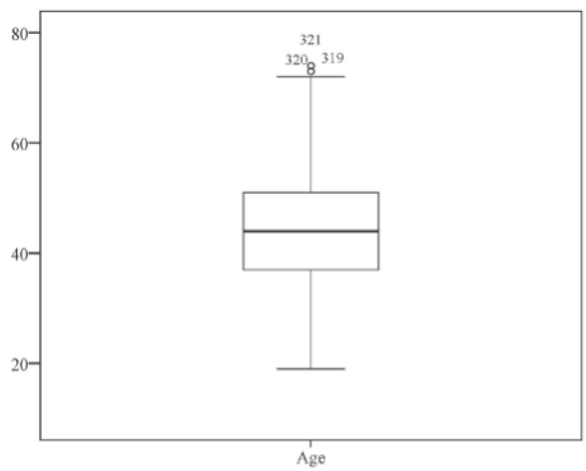
défini par les deux quartiles, inférieur (Q1) et supérieur (Q3). La ligne centrale qui la divise correspond à la médiane, dont la présentation peut être accentuée par des encoches (la moyenne peut aussi être projetée).

À l'extérieur de la boîte, les deux droites – les « moustaches » – définissent les limites inférieures et supérieures des valeurs « recevables ». Leur longueur ne doit pas dépasser une fois et demi la distance interquartile. Toute observation dont la valeur dépasse cette limite apparaît en point isolé, sur lequel il faut s'interroger.

Exemple du cas « Massif Vosgien » de la variable « ClasseAge » (tableau 1.5) : les trois valeurs extrêmes supérieures (2 points superposés) sont hors limites (319, 320, 321).

Tableau 1.5 – Diagrammes de dispersion d'une variable métrique

A - Diagramme en « boîte à moustaches »



B - Diagramme en tiges et feuilles

```
Age Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     2,00        1 .  99
    11,00        2 .  00012334444
    16,00        2 .  5556666777788899
    24,00        3 .  000001111222333333344444
    48,00        3 .  555555556666666666777777777788888888999999999999
    70,00        4 .  00000000000011111111111222222222222333333333333333444444444444444444
    59,00        4 .  55555555555566666666666666677777777778888888888999999999999
    42,00        5 .  000000000011111111111222222222333333333444
    28,00        5 .  5555555666666777777888999999
    10,00        6 .  0111222234
     6,00        6 .  667999
     2,00        7 .  02
     3,00 Extremes    (>=73)

 Stem width:        10
 Each leaf:       1 case(s)
```

Le diagramme en tige et feuilles (« stem and leaves diagram »)

Il permet d'analyser très rapidement la structure d'un échantillon de données en conservant les informations individuelles. À la différence d'un histogramme, auquel il s'apparente, ce diagramme n'écrase pas les données individuelles au sein des classes de valeurs qui ont été créées, mais propose une distribution des données en gardant trace de leur mesure initiale. Il nous renseigne donc sur la structure de distribution observée, à la fois de manière globale (elle est presque symétrique ici) et au sein de chaque classe, ce qui permet de juger de la cohérence des classes construites (nombre, intervalle de classe et bornes).

Une « tige » correspond à une *valeur* de la variable (la variable Âge du cas « Massif Vosgien » est ici décomposée en 12 classes, avec une classe extrême). Une « feuille » correspond à une *observation*.

> Dans l'exemple présenté, les 321 observations se répartissent entre les valeurs « 19 ans » (première classe) et « 72 ans » (douzième classe). Les « *tiges* » reprenant le premier nombre de la valeur de la classe, identifient les dizaines (par exemple, les âges dans la vingtaine – les deux tiges « 2 » – sont divisés en deux classes, de 20 à 24 ans pour l'une et de 25 à 29 ans pour l'autre). Les « *feuilles* » donnent les unités : elles représentent chaque observation dont les valeurs sont distribuées dans l'ordre croissant. Ainsi, la première « tige 2 » comprend trois observations qui ont 20 ans (trois « *feuilles* » avec 0), une seule qui a 21 ans et une autre qui a 22 ans (deux feuilles respectivement avec les unités 1 et 2) et quatre qui ont 24 ans (quatre feuilles avec l'unité 4) ; etc. Par ailleurs, ce type de disposition permet de voir si certaines valeurs sont plus fréquentes que d'autres ou d'identifier des séries (dans la cinquantaine – les deux « tiges 5 » – il y a peu de 54 et de 58 ans).

Toutes les valeurs de la variable sont ainsi réparties. Et la colonne de gauche « effectif » donne le nombre d'observations par classe créée. Ainsi, la première classe 2 contient onze observations correspondant aux valeurs comprises entre 20 et 24 ans, la première classe 4 contient 70 observations courant de 40 ans (12 éléments) à 44 ans (18 éléments), etc. Quant à la classe « extrêmes », elle contient trois observations égales ou supérieures à 73 ans. Ce diagramme renseigne donc sur la structure de distribution observée, de manière globale (elle est symétrique ici) et au sein de chaque classe, ce qui permet de juger de la cohérence des classes construites (nombre, intervalle de classe et bornes).

Outre les renseignements fournis sur la structure de la distribution, ces deux types de diagramme permettent d'effectuer un contrôle de la qualité des données recueillies, particulièrement en ce qui concerne les valeurs extrêmes. Ils permettent de détecter des données incohérentes ou hors normes (par exemple, la présence d'une valeur qui n'est pas incluse dans la métrique de la variable) et de vérifier la qualité des informations à traiter.

© Dunod – Toute reproduction non autorisée est un délit.

3 Les indicateurs statistiques descriptifs d'une variable

Malgré la forte valeur de communication des représentations graphiques, elles ne sauraient pouvoir se substituer aux indicateurs numériques. En effet ces derniers fournissent l'information chiffrée précise nécessaire pour analyser les variables et leurs relations, et pour justifier leur emploi dans certaines approches statistiques plus complexes. Le tableau 1.6 présente les principaux indices statistiques selon leur objectif (tendance centrale, dispersion et forme) et la nature métrique ou non des variables. Le tableau 1.7 illustre ces statistiques appliquées sur des variables du fichier de données relatif au cas « Massif Vosgien ».

Tableau 1.6 – Les indices statistiques selon la nature des variables

Type de variable		INDICES STATISTIQUES		
Nature	Métrique	Valeur centrale	Dispersion	Forme
Non métrique	nominale	mode	fréquence	-
	ordinale	mode médiane	étendue fréquence fractile	-
Métrique	intervalle ou rapport	mode médiane moyenne	étendue / fractile variance écart-type	asymétrie (*skewness*) aplatissement (Kurtosis)

Tableau 1.7 – Statistiques pour cinq variables du tableau 1.1 (sous Statgraphics)

	DurSej	AdapteVacances	SatisAccueil	SatisNeige	Âge
Effectif	321	321	321	321	321
Moyenne	4,234	3,654	3,302	3,433	44,022
Écart-type	0,955	1,065	1,072	0,768	10,577
Coefficient. de variation	22,55 %	29,13 %	32,46 %	22,37 %	24,03 %
Minimum	1,0	1,0	0	1,0	19,0
Maximum	7,0	5,0	6,0	6,0	74,0
Étendue	6,0	4,0	6,0	5,0	55,0
Asymétrie brute	-0,338	-0,776	-0,075	0,039	1,119
Asymétrie standard.	-2,471	-5,674	-0,545	0,287	0,867
Aplatissement brut	0,090	0,098	-1,105	0,296	0,192
Aplatissement std.	0,329	0,358	-0,384	1,083	0,701

3.1 Les statistiques de tendance centrale (tableau 1.8)

Elles servent à *résumer* une série d'observations avec une seule valeur indiquant le « centre » des données.

– Le *mode* [« Mo »] (« mode »)[1] indique la modalité de la variable qui cumule le plus d'observations et montre la tendance générale de la distribution. La variable peut être composée d'un seul ou de plusieurs modes, visibles avec son histogramme. La présence de plusieurs *classes modales* indique alors que la distribution n'est pas homogène et qu'il y a des sous-groupes d'observations à différencier. Le mode est la seule statistique de tendance centrale pour résumer une variable nominale.

La v2-TypeGroupe (*cf.* figure 1.2) a deux classes modales : les modalités 4 et 5 comportent respectivement 24,9 et 25,2 % des observations.

Tableau 1.8 – Les indices de tendance centrale

Mode :	Modalité ou valeur de la variable recueillant le plus d'observations
Médiane :	Modalité ou valeur de la variable divisant l'ensemble des observations en deux groupes de taille égale
Moyenne :	Valeur résumant l'ensemble des mesures d'une variable métrique. Pour les données brutes, elle est calculée avec la formule : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ Mais pour les variables regroupées en k modalités (classes), elle est calculée avec la formule : $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$ Quand la variable est regroupée en classes, ce sont les centres de classes qui sont utilisés à la place des x_i

– La *médiane* [« Me »] (« median ») indique la valeur de la variable qui partage l'ensemble des observations en deux groupes égaux. La médiane et le mode sont les deux seules statistiques de tendance centrale utilisables pour résumer une variable ordinale.

La 161e observation (le nombre entier supérieur à 321/2 = 360,5) correspond à la modalité 4 de la variable « TypeGroupe ».

– La *moyenne arithmétique* [« m » ou « $\bar{x}$ »] (« *mean* » ou « *average* ») calcule la valeur centrale exacte de la distribution étudiée et résume l'ensemble des données.

La moyenne d'appréciation de l'accueil pendant le séjour vaut 3,3 (sur 5) (SatisAccueil) et la moyenne d'âge des vacanciers se situe à 44 ans.

1. Nous donnons entre […] le symbole usuel et entre (…) les termes équivalents anglais tels qu'ils apparaissent dans les logiciels anglo-saxons quand ils diffèrent du français.

© Dunod – Toute reproduction non autorisée est un délit.

Mode, médiane et moyenne sont calculables pour toutes les données métriques : échelles d'intervalle et de rapport.

Remarque

Une variable non métrique *binaire*, mais prenant les valeurs 0 et 1 est considérée comme métrique. S'agissant d'une variable booléenne, on peut calculer sa moyenne qui correspond à la fréquence : ce pourrait être le cas des variables 4 et 5 (pratique ski de fond et alpin) si leurs modalités étaient 0/1 au lieu de 1 & 2.

3.2 Les statistiques de dispersion (tableau 1.9)

Les statistiques de tendance centrale ne renseignent pas sur la structure de la distribution et la dispersion des observations autour de cette valeur centrale : c'est le rôle des statistiques de dispersion. Leur mesure permet de décider du degré d'homogénéité ou, au contraire, d'hétérogénéité de la distribution des observations.

Tableau 1.9 – Les indices de dispersion

Fréquence : f	proportion des observations présentant une modalité de la variable : ($f = n / N$)
Étendue : ω	écart entre les valeurs extrêmes observées sur la variable : $\omega = x_{i\,max} - x_{i\,min}$
Fractile	valeurs partageant la distribution des observations en un certain nombre d'intervalles égaux : les quartiles la divisent en 4 parties égales contenant 25 % des observations ; les déciles la divisent en 10 parties et les centiles en 100 parties. D'autres valeurs de fractiles peuvent être choisies.
Variance : σ^2:	Moyenne des carrés des écarts de chaque observation à la moyenne des observations ; elle informe sur l'étendue de la variation des observations autour de la moyenne. Si le calcul se fait à partir des données brutes la variance est calculée avec la formule : $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ mais si calcul se fait à partir des données regroupées en k modalités (classes), la variance est calculée avec la formule suivante : $\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2$ Les logiciels statistiques font la différence entre la variance d'une population et la variance d'un échantillon. La première est obtenue en divisant la somme des carrés des écarts de chaque observation à la moyenne des observations par n, alors qu'elle l'est par n-1 pour la variance d'un échantillon.
Écart-type : σ	Même information que la variance mais exprimée dans la métrique de la variable. Elle se calcule si les données sont brutes : $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ mais avec cette formule si la variable a été regroupée en modalités ou classes : $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}$

– La *fréquence* [« *f* »] (« relative frequency »)[1] renseigne sur la dispersion des observations entre les modalités ou les valeurs d'une variable. Quand toutes les modalités ont la même fréquence, la dispersion est maximale :

> L'exemple de la variable « 2-TypeGroupe » (figure 1.1) : les fréquences sont voisines pour quatre des cinq modalités : respectivement, 19,9 % - 9,98 % -19,9 % -24,9 % et 25,2 %.

La fréquence est aussi une statistique utile pour effectuer des comparaisons, car elle permet de comparer la dispersion des données entre des échantillons d'observations de taille différente.

– Les *fractiles* définissent la valeur qui partage un échantillon d'observations en un certain nombre de classes d'effectif égal. Quartiles, déciles et percentiles sont les plus usuels et leur choix dans une analyse dépend des objectifs de l'étude et de la taille des échantillons. Les fractiles permettent d'avoir une juste appréciation de la réalité de la dispersion des observations et compensent les limites de l'étendue [« ω »] (« range ») qui est très dépendante d'observations marginales (puisqu'elle mesure l'écart entre les valeurs extrêmes).
– Les *déciles* partagent l'ensemble des observations recueillies en dix parties égales et les percentiles, en cent parties égales.
– Les *quartiles* [Q1, Q2 et Q3] partagent ce même ensemble en quatre parties égales contenant chacune 25 % de l'effectif total. Le second quartile correspond à la médiane [Q2 : 50 % des observations avant et 50 % après].
– L'*intervalle interquartile* (« interquartile range ») est la différence entre les quartiles extrêmes Q3- Q1 qui contient donc 50 % des observations, laissant respectivement 25 % à droite et 25 % à gauche. Pour la variable « Âge » ces valeurs sont respectivement : Q1 = 37 ans, Q2 = 44 ans, Q3 = 51 ans. Les quartiles peuvent être utilisés afin de déterminer deux groupes d'observations opposés pour comparer différentes variables.

> Pour savoir si une différence d'âge entraîne des pratiques sportives différentes (ski de fond ou alpin dans le cas étudié), il suffit de comparer sur les seules classes générées par les deux quartiles extrêmes (Q.1 et Q.3) de la variable « âge » le taux de pratique des plus jeunes et des plus âgés. Cette comparaison montre que le niveau de pratique est identique pour le ski de fond et le ski alpin chez les plus jeunes (< 38 ans : Q.1) comme chez les plus âgés (> 50 ans : Q.3).

– La *variance* [σ^2] et l'*écart-type* [σ][2] (« standard deviation ») renseignent sur l'étendue de la dispersion des observations d'une variable métrique autour de sa moyenne ; c'est une mesure de la distance moyenne des observations à la moyenne arithmétique de la variable étudiée.
Ces deux statistiques, dont la fonction est identique, ne diffèrent que par leur métrique : la variance est calculée sur la base du carré des écarts de chaque obser-

1. En anglais, il importe de ne pas confondre, *frequency* qui correspond à l'opération de dénombrement (l'effectif), avec *relative frequency* qui correspond à la fréquence.
2. On lui donne aussi à l'écart-type le nom d'écart quadratique moyen puisque c'est la moyenne d'ordre 2 (ou moyenne quadratique) des écarts à la moyenne arithmétique.

© Dunod – Toute reproduction non autorisée est un délit.

vation à la moyenne de la variable ; l'écart-type étant sa racine carrée, il mesure donc la dispersion dans la même unité métrique que la variable originelle (qui est aussi celle de la moyenne). Lorsque les données ont été standardisées (de moyenne 0 et d'écart-type 1), on peut donc exprimer la distance d'une observation quelconque à sa moyenne en fraction d'écart-type et apprécier ainsi son caractère « normal » ou son caractère « marginal ».

– Le *coefficient de variation* [« CV »] correspond au rapport de l'écart-type à la moyenne [$CV = \sigma/\bar{x}$]. Indépendant des unités de mesure, il permet de comparer la dispersion de distributions différentes (par exemple la distribution du revenu moyen dans des pays de monnaie différente ou celle issue d'échelles d'opinion en 5 et en 11 points). Multiplié par 100, il exprime un pourcentage de variation.

3.2 Les statistiques de forme (tableau 1.10)

D'un certain point de vue, les coefficients qui renseignent sur la forme d'une distribution peuvent être considérés comme des indices de dispersion, puisque la forme de la distribution dépend de la structure de dispersion des données. Ils permettent de vérifier si la distribution peut être approximée par une loi normale (*cf.* 4.2) ; ce contrôle est important puisque nombre d'applications statistiques ne sont justifiées que si les variables étudiées se distribuent normalement.

Tableau 1.10 – Les indices de forme

Asymétrie ou « *Skewness* »	Mesure de l'écart entre la moyenne et la médiane qui permet de caractériser le degré de symétrie de la distribution par rapport à la valeur centrale $Skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns_x^3}$
Aplatissement ou « Kurtosis »	Mesure le niveau de concentration des observations en regard d'une distribution normale $\text{Kurtosis} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns_x^4}$

– Le *coefficient d'asymétrie* [« *skewness* »] renseigne sur l'équilibre des observations autour de la moyenne. La valeur du *skewness* est nulle si les observations se répartissent équitablement autour de la moyenne. Si les observations sont décalées vers les plus petites valeurs (vers la gauche) de la distribution, ce coefficient est positif : la moyenne est supérieure à la médiane ; si elles sont plutôt décalées vers la droite (valeurs de la distribution les plus élevées), le *skewness* est négatif : la moyenne est inférieure à la médiane.

– Le *coefficient d'aplatissement* [« Kurtosis »][1] renseigne sur le degré de concentration des observations autour de la moyenne : soit la courbe est pointue (forte

1. Le classique coefficient d'aplatissement est quelquefois dénommé « coefficient de concentration » – *cf.* Y. Evrard et *alii.*, 2000, p. 335 ; cette appellation peut induire la confusion avec *l'indice de concentration* qui est une mesure

concentration ou distribution *leptokurtique*) : dans ce cas, sa valeur est positive (ou supérieure à 3) ; soit elle est étale ou étirée (faible concentration ou distribution *platikurtique*) : alors sa valeur est négative (ou inférieure à 3).

Remarque importante

Dans une loi normale, le coefficient d'aplatissement vaut 3 avec la formule originelle de Pearson. Mais, pour conserver cette loi normale comme référence, nombre de logiciels anglo-saxons retranchent 3 à sa valeur théorique pour avoir un Kurtosis de référence à 0 (formule de Fisher). C'est la raison pour laquelle, compte tenu de cette pratique très répandue, la variation de la valeur du Kurtosis de Pearson où la valeur 3 est la norme, est précisée ci-dessus. De ce fait, la valeur de ce coefficient peut varier d'un logiciel à l'autre en fonction de la formule utilisée – qu'il faut donc vérifier – pour obtenir une estimation non biaisée.

Compte tenu de l'usage, il est admis que dans une distribution normale, ces coefficients sont nuls (*cf.* remarque ci-dessus). En conséquence, on admet dans la pratique qu'une distribution est normale si l'indice de symétrie est inférieur à 1 et celui d'aplatissement inférieur à 1,5. Il existe une version standardisée de ces deux coefficients (logiciel Statgraphics) qui exprime leur intervalle de variation en écart-type : si la valeur de chacun de ces coefficients est comprise entre ± 2 écarts-types, on admet que symétrie et aplatissement sont compatibles avec les spécifications d'une loi normale. En l'absence de normalité, une transformation de la variable, de type log^x par exemple, est à envisager.

Dans le tableau 1.7, les variables « SatisAccueil », « SatisNeige » et « Âge » se distribuent normalement puisque les valeurs de leurs coefficients d'asymétrie et d'aplatissement sont faibles et comprises dans l'intervalle de $\pm 2\sigma$. Mais il n'en est pas de même pour la durée du séjour (« DurSéj ») et l'adaptation du Massif aux vacances (« AdaptéVacances ») : leur asymétrie est élevée et décalée à gauche (signe négatif) vers les petites valeurs. Il faudra donc les transformer si leur intégration dans un modèle statistique implique la normalité de leur distribution.

3.3 Les statistiques de liaison (tableau 1.11)

Une dernière catégorie de statistiques permet de mesurer le lien qui existe entre paires de variables. Les statistiques de liaison renseignent sur l'intensité et le sens du lien qui unit deux variables :

- entre variables métriques, cette statistique mesure la force d'une *liaison linéaire* ;
- entre variables ordinales, elle exprime le degré de similitude entre deux classements ;
- entre variables nominales, elle traduit le niveau d'association entre les deux variables.

de la dispersion relative utilisée pour décrire la distribution statistique de variables particulières, avec lequel il ne doit pas être confondu – *cf.* B. Grais, 1979, pp. 167-171.

© Dunod – Toute reproduction non autorisée est un délit.

La plupart de ces indices sont normés (mais pas tous) et varient dans l'intervalle [– 1 ; 0 ; + 1] : la valeur 0 indique l'absence de lien, donc l'indépendance entre les variables ; et la valeur 1 traduit une association maximale. Une valeur positive du coefficient indique que les deux variables évoluent de manière parallèle ; une valeur négative, qu'elles évoluent en sens inverse : quand les valeurs de l'une croissent, celles de l'autre décroissent. Comme ces coefficients sont sensibles aux effets de taille, tel que le *coefficient de corrélation* de Bravais – Pearson, il est nécessaire de tester leur significativité pour vérifier la réalité du lien.

Tableau 1.11 – Les indices de liaison

Nature	Échelle	Nominale binaire	Nominale non binaire	Ordinale	Métrique
Non métrique	**nominale binaire**	Coefficients : – Phi (Φ) – Q de Yule			Coefficient point bisérial
	nominale non binaire		Coefficients : – de contingence C – T de Tschuprow – V de Cramer – Lamda (λ) de Goodman et Kruskal		
	ordinale			Coefficients de rangs : – rhô de Spearmann (ρ) – tau de Kendall (τ) – gamma de Goodman et Kruskal (γ)	Coefficient point multibisérial
Métrique	**intervalle ou rapport**	Coefficient point bisérial		Coefficient point multibisérial	Coefficient de corrélation de Bravais – Pearson

4 Vérifier la normalité d'une variable métrique

Pour les utiliser, nombre de méthodes statistiques impliquent que la distribution des variables soit statistiquement normale. Aussi, diverses approches, probabilistes ou empiriques, permettent de le vérifier.

4.1 Les approches empiriques et graphiques

– L'*histogramme de distribution* donne une idée du degré de normalité de la distribution (ou de la loi que suit cette distribution) en le superposant à celui de la loi normale ayant même moyenne et même variance. Mais cette pratique ne permet pas de tester avec exactitude la normalité de la distribution observée.

– La *boîte à moustaches (box-plot)* permet de visualiser rapidement la symétrie de la distribution des données réelles et la présence de valeurs atypiques.
– Le *diagramme quantile-quantile (Q-Q plot)* permet de comparer deux échantillons de taille différente. Cela consiste à comparer la distribution observée avec une distribution normale en représentant les quantiles de la distribution observée par rapport aux quantiles de la distribution normale. Le « Q-Q plot » est un nuage de points où les quantiles observés sont placés en abscisse et les quantiles théoriques de la loi normale, en ordonnée. Si la variable observée correspond à une distribution normale, les points sont concentrés de part et d'autre d'une droite.

Les tests empiriques de normalité sont très faciles à mettre en œuvre, mais ils manquent de rigueur statistique. On fait alors appel à des approches probabilistes.

4.2 Les approches probabilistes

Il existe diverses techniques probabilistes pour vérifier si une distribution est ou n'est pas statistiquement normale : la plupart des logiciels de statistique permettent de calculer certains tests pour évaluer la normalité (tests de Kolmogorov-Smirnov, de Shapiro-Wilk, d'Andersen-Darling, de Lilliefors, de Rayan-Joiner, de Jarque-Bera, etc.). Toutes ces approches vérifient l'adéquation d'une distribution à la loi normale. La plupart des tests probabilistes sont des variantes du test de Kolmogorov-Smirnov reposant sur les coefficients d'asymétrie et d'aplatissement.

Deux tests sont particulièrement intéressants, l'un pour sa facilité de mise en œuvre et l'autre, bien adapté aux petits échantillons, pour sa puissance.

– Le **Test de Jarque-Bera** : simple à calculer, il vérifie si une distribution suit une loi normale. Utilisable sur de petits échantillons, il repose sur les coefficients d'asymétrie (*skewness*) et d'aplatissement (Kurtosis) et consiste à calculer la statistique suivante si elle repose sur le calcul originel de Pearson (norme fixée à 3) :

$$JB = n\left[\frac{Skewness^2}{6} + \frac{(Kurtosis - 3)^2}{24}\right]$$

ou sur cette autre statistique, si elle repose sur la transformation de Fisher qui fixe la norme à 0 (donc soustraction de 3) :

$$JB = n\left[\frac{Skewness^2}{6} + \frac{(Kurtosis)^2}{24}\right]$$

La normalité est testée de manière indirecte en vérifiant, non pas si les données suivent une loi normale, mais si les indices de forme (*skewness* et *Kurtosis*) de la distribution testée sont identiques à ceux d'une loi normale de même moyenne et de même variance. Pour de grands échantillons, la statistique de Jarque-Bera suit une

© Dunod – Toute reproduction non autorisée est un délit.

loi du Khi-deux à 2 degrés de liberté. Pour des petits échantillons, Bera et Jarque[1] proposent des valeurs critiques à ne pas dépasser pour des seuils de risque de 0,1 et de 0,05 (tableau 1.12). Si la valeur calculée dépasse la valeur théorique de la table, on rejette l'hypothèse de normalité, sinon on l'accepte.

Tableau 1.12 – Valeurs critiques de la statistique de Jarque-Bera

Taille de l'échantillon	$\alpha = 0,1$	$\alpha = 0,05$
20	2,13	3,26
30	2,49	3,71
40	2,70	3,99
50	2,90	4,26
75	3,09	4,27
100	3,14	4,29
125	3,31	4,34
150	3,43	4,39
200	3,48	4,43
250	3,54	4,51
300	3,68	4,60
400	3,76	4,74
500	3,91	4,82
800	4,32	5,46
∞	6,41	5,99

Table extraite de NEWBOLD, P., CARLSON, W. L. & THORNE, B. (2010) Statistics for Business and Economics, 7th edition, Pearson, p. 647.

À partir des données brutes (et dont la norme est 0) fournies dans le tableau 1.6, la statistique de Jarque-Bera prend la valeur suivante pour la variable « Durée du séjour » :

$$JB_{obs} = n\left[\frac{Skewness^2}{6} + \frac{(Kurtosis)^2}{24}\right] = 321\left[\frac{0,338^2}{6} + \frac{(0,090)^2}{24}\right] = 6,22$$

Les valeurs théoriques de cette statistique pour une taille d'échantillon de 321 et des seuils d'erreur de 0,1 et 0,05 sont :

$$JB_{th\alpha=0,1} = 3,68 \; et \; JB_{th\alpha=0,05} = 4,60$$

La valeur observée, bien supérieure à la valeur théorique de la table, oblige à conclure que la variable n'est pas distribuée normalement, avec un niveau de confiance de 99 %.

1. Bera, A.K. and Jarque, C.M. (1981): "An efficient large-sample test for normality of observations and regression residuals", Australian National University Working Paper in Economics and Econometrics 40.

Pour la variable « SatisNeige », la valeur du test valant 1,25 (< 4,60) cela conduit à conclure que cette variable se distribue normalement.

$$JB_{obs} = n\left[\frac{Skewness^2}{6} + \frac{(Kurtosis)^2}{24}\right] = 50\left[\frac{0,039^2}{6} + \frac{(0,296)^2}{24}\right] = 1,25$$

– Le **Test de Shapiro-Wilk :** c'est le test probabiliste le plus fiable appliqué sur de petits échantillons et il repose sur le calcul de la statistique suivante :

$$W = \frac{\sum_{i=1}^{[n/2]} a_i \left(x_{n-i+1} - x_{(i)}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

où x_i représente le i^{ème} plus petit nombre dans l'échantillon ; $[n/2]$ est la partie entière du rapport n/2 ; a_i sont des constantes générées à partir de la moyenne et de la matrice de variance-covariance des quantiles d'un échantillon de taille n suivant la loi normale. L'hypothèse nulle postule que la distribution est normale et vérifiée si la probabilité critique du test est élevée. Certes, ce test nécessite des calculs un peu compliqués (et oblige de plus à se référer à deux tables de valeurs critiques), mais les logiciels offrant ce test (la plupart) se chargent des calculs.

Le tableau 1.13 donne, pour les variables « Durée du séjour » et « SatisNeige », les résultats des tests de Lilliefors et de Shapiro-Wilk, calculés avec SPSS pour un seuil de confiance de 95 %. Ces résultats mènent à rejeter l'hypothèse nulle de normalité pour les deux variables (ce qui est en contradiction avec le test de Jarque Bera pour la variable « SatisNeige »). Le test de Shapiro-Wilk étant le plus fiable, c'est sur son résultat qu'il faut se reposer : la normalité des deux variables n'est pas respectée.

Tableau 1.13 – Résultats des tests de Kolmogorov-Smirnov et de Shapiro-Wilk par SPSS

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistiques	ddl	Sig.	Statistiques	ddl	Sig.
Durée du séjour	,208	321	,000	,900	321	,000
SatisNeige	,262	321	,000	,857	321	,000

a. Correction de signification de Lilliefors

© Dunod – Toute reproduction non autorisée est un délit.

L'ESSENTIEL

■ *Construire les variables à partir des données extraites des fichiers étudiés*

Une décision managériale s'appuie sur des informations extraites de fichiers de données qui doivent être pertinentes, fiables et de bonne qualité. La construction des variables à partir de ces données initiales est une étape cruciale lors du traitement statistique. La construction et la caractérisation des variables impliquent plusieurs étapes aussi importantes les unes que les autres :

– Transformer les données collectées en variables sous un format exploitable par les logiciels de traitement statistique. Cette étape implique de sélectionner les données pertinentes pour répondre à la question managériale posée et de transformer ces données en variables.

– S'assurer de la qualité, de l'exactitude et de la pertinence des données : détecter les différentes sources d'erreurs, remplacer les données manquantes, sélectionner les variables pertinentes.

■ *Décrire, caractériser et analyser les variables construites*

Analyser la distribution statistique et caractériser chaque variable avec les paramètres descriptifs adaptés à l'information recherchée en fonction de leur nature, métrique ou non-métrique : mesure des tendances centrales, caractéristiques de dispersion et de forme ; puis les représenter visuellement avec les graphiques correspondant. Enfin, vérifier la normalité de leur distribution quand le problème étudié l'exige.

Chapitre 2

Comment comparer des échantillons

OBJECTIFS

- Connaître les différentes classes de comparaison pour identifier celle qui correspond à la problématique décisionnelle rencontrée et construire le plan d'observation adapté.
- Compte tenu de la nature des données disponibles, choisir et utiliser l'outil statistique le mieux adapté à la décision pour opérer la comparaison voulue et formuler les hypothèses à valider.
- Passer du résultat statistique observé (acceptation ou rejet de l'hypothèse nulle) à la décision opérationnelle qui en découle.

SOMMAIRE

SECTION 1 Méthodologie de la comparaison et conditions d'application

SECTION 2 Passer du résultat statistique à la décision managériale : traitement du cas « Cirage »

Nombre de décisions opérationnelles reposent sur la comparaison de résultats liés à l'activité de l'entreprise ou obtenus par enquête. Cette comparaison permet de vérifier si l'écart constaté entre les valeurs comparées est significatif d'une différence. Il existe un certain nombre de tests statistiques – les tests *d'ajustement* et *d'inférence* – qui permettent d'évaluer objectivement ces écarts. Cependant, la question se pose de savoir quel type de test utiliser dans un contexte décisionnel donné.

Aussi, le traitement d'une problématique décisionnelle fondée sur un test de comparaison statistique – de sa formulation à sa résolution – est-il facilité en respectant les trois étapes proposées ci-dessous pour conduire ce test :

1. *Identifier le contexte décisionnel de la comparaison*
   – Expliciter la problématique managériale et le cadre décisionnel
   – Identifier la comparaison statistique adaptée au problème décisionnel
2. *Mettre en œuvre les procédures statistiques adaptées*
   – Choisir la méthode et l'outil statistiques pertinents
   – Formuler les hypothèses statistiques
3. *Passer du résultat statistique à la prise de décision*
   – Analyser les données et interpréter les résultats statistiques
   – Formuler les conclusions managériales qui en découlent.

Section 1 MÉTHODOLOGIE DE LA COMPARAISON ET CONDITIONS D'APPLICATION

Dans le cadre de la politique de relance de l'un de ses produits, une petite entreprise envisage de modifier la présentation commerciale de l'une de ses crèmes de cirage exclusivement vendue en boîte en proposant un conditionnement en tube (avec un type d'applicateur assez novateur). Le tube étant globalement perçu par la clientèle comme plus « noble » que la boîte, ce choix pourrait renforcer l'image de qualité du produit et le rendre encore plus concurrentiel.

Pour décider de la pertinence de cette substitution – ce nouveau type de tube entraînant un léger surcoût de production – le chef de projet réalise un test d'accueil (à domicile en deux visites) auprès d'un échantillon de 420 adultes censés représenter la population des utilisateurs réguliers de cirage. Il espère ainsi obtenir les informations nécessaires pour répondre aux différentes questions qu'il se pose avant d'arrêter sa décision finale.

Cas d'illustration

Cirage

Décision 1

Pour vérifier si le tube [T] est favorablement perçu chez les plus modestes des utilisateurs actuels et réguliers du cirage en boite [B], le chef de produit teste simultanément les deux conditionnements comportant la même pâte auprès d'un sous-échantillon composé de 200 personnes appartenant effectivement aux classes les plus modestes [M]. Les résultats sont les suivant : 90 personnes préfèrent la boîte [B] et 110 personnes préfèrent le tube [T]. Au vu de ces résultats, le responsable du projet s'interroge sur le risque commercial que l'entreprise prendrait à remplacer la boîte par le tube auprès de cette cible.

Décision 2

Parallèlement aux préoccupations sur le choix du conditionnement, le chef de produit et le responsable de production se préoccupent de définir la bonne consistance de la pâte en tube par rapport aux attentes de la clientèle. Les résultats suivants concernent l'appréciation de la fluidité chez les utilisateurs de boîte pour lesquels l'incertitude était la plus grande. Cette mesure a été effectuée à l'aide d'une échelle verbale préalablement étalonnée en 5 niveaux (codifiés de 1 à 5) dont le tableau ci-dessous présente les résultats :

«Après usage, diriez-vous que le cirage - crème utilisé vous semble…» :

	x	n	Résultats
– beaucoup trop fluide	1	30	Score moyen : 2,45 Écart type : 0,867 Effectif : 200
– un peu trop fluide	2	70	
– juste bien	3	80	
– un peu trop dure	4	20	
– beaucoup trop dure	5	0	

Au vu de ce résultat (2,45), le chef de produit et le responsable de la production ne sont pas d'accord sur l'intérêt qu'il y aurait ou non à modifier la consistance de la pâte. L'un considère que « 2,45 c'est presque 3 (juste bien) et qu'il est inutile de modifier la consistance » et l'autre, plus circonspect, se demande s'il ne conviendrait pas de la modifier pour « bien coller aux attentes ».

Décision 3

L'ensemble des résultats précédents interroge le responsable du projet qui se demande s'il ne faudrait pas segmenter les clientèles pour proposer à chaque cible le conditionnement le plus adéquat : conserver la boîte chez les plus modestes [M] et imposer le tube chez les plus aisés [A]. Pour éclairer son jugement, il fait réaliser un questionnaire complémentaire pour tester la préférence entre le tube et la boîte auprès des 220 personnes appartenant aux classes plus aisées [A]. Comme 90 personnes préfèrent la boîte et 130 le tube, il se demande s'il est opportun de réserver le tube au seul segment de la clientèle la plus aisée [A].

☞

© Dunod – Toute reproduction non autorisée est un délit.

Décision 4

Compte tenu des caractéristiques du nouvel applicateur (T), le chef de projet s'est demandé si les habitudes d'utilisation du cirage pouvaient avoir une incidence sur la perception des performances de cet applicateur. Toutes les personnes de l'échantillon étudié ont été réparties en deux groupes sur la base de leurs habitudes d'usage : l'un composé de gros (260) et l'autre de petits (160) utilisateurs de crème cirage. Sur la base des résultats obtenus sur plusieurs indicateurs métriques, un score composite de performance (note comprise entre 0 et 10) a été calculé.

	Petits utilisateurs [Pu]	Gros utilisateurs [Gu]
Moyenne ($\bar{x}$)	5,7	6,3
Écart-type (s)	1,1	1,2
Effectifs (n)	160	260

Le chef de produit pense que ces résultats confirment son intuition, à savoir que le niveau de performance du nouvel applicateur est perçu de manière pratiquement identique quel que soit le niveau d'utilisation du cirage.

Décision 5

Il s'est aussi inquiété de savoir si cet applicateur présentait une réelle valeur concurrentielle. Pour ce faire, un tri a permis d'extraire sur l'échantillon les 120 personnes utilisatrices d'applicateurs à l'exclusion de tout autre type de conditionnement. Lors de la première visite, une question portait sur la satisfaction générale des répondants à l'égard des produits cirant utilisés [U] (note globale comprise entre 0 et 5) ; au terme de la seconde visite, cette même question était de nouveau posée pour le seul tube testé avec son applicateur [T]. Compte tenu des résultats observés, le chef de produit s'interroge sur la réalité de l'avantage concurrentiel de son applicateur :

Applicateur :	Usuel [U]	Tube [T]
Moyenne ($\bar{x}$)	2,95	3,22
Écart-type (s)	1,25	1,32
Effectifs (n)	120	120

Décision 6

Le dépouillement du questionnaire a aussi permis de constater que sur le groupe des personnes aisées [A], il y avait une certaine convergence dans les déclarations recueillies sur l'importance reconnue des soins qu'il convenait d'apporter aux chaussures, quel que soit le sexe. Cette convergence ne semble pas caractériser l'autre groupe [M : « les modestes »] dont la dispersion semble plus grande et, en apparence, variable selon le sexe. Comme cette variable peut avoir une incidence sur le choix du mode d'entretien et l'achat du type de cirage, il a été décidé de vérifier cet aspect sur ce seul segment de clientèle dont le tableau suivant présente les résultats :

☞

Est-ce très important, juste important ou pas du tout important d'apporter beaucoup de soin à ses chaussures ?

sur segment [M]	*x*	Hommes	Femmes
Très important…	*1*	20 %	40 %
Juste important …	*2*	60 %	30 %
Pas du tout important…	*3*	20 %	30 %
	Effectif :	100	100

Décision 7

En prévision de la campagne publicitaire pour annoncer son nouveau conditionnement – si la décision de lancement est prise – l'entreprise hésite entre trois argumentaires. Elle souhaite utiliser celui qui valoriserait le mieux son offre face à la concurrence. Elle teste ces trois argumentaires sur la cible des seuls utilisateurs d'une quelconque marque concurrente (soit 331 personnes ; mais chacune n'est exposée qu'à un seul argumentaire) pour mesurer la crédibilité et l'incitation à l'essai. Ces deux critères sont fusionnés dans un indicateur métrique et chaque argumentaire est soumis à une centaine de personnes. Les résultats sont les suivants :

Argumentaire	Effectifs	Moyenne	Intervalle de confiance de la moyenne	Écart-type
1	98	3,65	[3,558 ; 3,748]	0,943
2	124	6,50	[6,401 ; 6,599]	1,108
3	109	5,91	[5,763 ; 6,053]	1,512

1 Identifier le contexte décisionnel pour choisir la méthode de comparaison pertinente

Comparer consiste à mesurer l'importance des écarts constatés entre des échantillons évalués sur un même critère. Il faut donc commencer par préciser si la décision managériale envisagée dépend d'une absence d'écart ou, au contraire, d'un écart vraiment significatif : c'est la base de l'hypothèse statistique à valider.

Remarque

La décision finale peut aussi résulter de plusieurs hypothèses à considérer successivement, à l'exemple du cas « Cirage ». Au terme du test d'accueil réalisé, le chef de projet espère que les clients ciblés percevront une différence de qualité significative en passant de la boîte au tube, mais que, au minimum, ils ne percevront pas de différence entre son produit conditionné en tube et l'offre du concurrent leader, au point de les confondre (*absence de différence statistiquement significative souhaitée face au concurrent*), si ce n'est de le dépasser (*qualité de son produit perçue comme supérieure*). Si ces conditions sont réunies, il peut considérer que sa nouvelle offre est une alternative crédible au produit concurrent et décider de la mettre sur le marché.

© Dunod – Toute reproduction non autorisée est un délit.

Aussi, pour décider du test de comparaison à utiliser pour valider cette hypothèse statistique, il faut d'abord considérer simultanément :
– la nature – métrique ou non – des données sur laquelle s'appuie la décision ;
– le nombre des échantillons de mesures à comparer et leur caractère indépendant ou apparié.

1.1 La nature des données exploitées

Il est possible de *comparer* des valeurs ponctuelles – limité ici aux moyennes et aux fréquences – ou des distributions de valeurs. On peut aussi vouloir *vérifier* si la distribution d'une variable mesurée sur un échantillon est conforme à celle d'une population connue.

Pour comparer des moyennes, les variables doivent impérativement être métriques (échelles d'intervalle ou de rapport) ; les variables non métriques (nominales ou ordinales) ne permettent de comparer que des fréquences ou des médianes liées aux classes ou aux modalités de la variable (*cf.* ch. 1).

> **Quelques exemples**. Comparer le revenu moyen des actifs de 25-50 ans à celui des 50-65 ans (moyennes) ou le taux de multi-équipement en smartphones des 18-35 ans à celui des 35-50 ans (fréquences). Vérifier la qualité d'un échantillon construit (sur la base de quotas) avec la population qu'il est censé représenter (comparaison de distributions).

Comparer des valeurs ponctuelles (moyennes et fréquences) ou comparer des distributions implique d'utiliser des tests différents.

– les *tests de comparaison* simple, cas le plus usuel, où les paramètres statistiques mesurés sont comparés ; par exemple : comparer le nombre de demandes d'information pour un nouveau service bancaire selon le niveau d'utilisation d'un panier de services existant déjà ;
– les *tests d'ajustement* où une distribution observée est comparée à une distribution connue, à l'exemple de la vérification de la représentativité d'un échantillon aléatoire par rapport à une population de référence.

Cependant, la nature métrique ou non métrique des variables étudiées oriente vers une classe d'outils spécifiques :
– Pour une *variable métrique* dont la distribution des valeurs suit une loi normale (loi de Laplace-Gauss), on utilise un test paramétrique de comparaison de moyennes (principalement le *test en Z* et le test *t de Student*).
– Pour une *variable* qui n'est *pas métrique* : on utilise un test non paramétrique pour comparer soit des fréquences, soit des médianes (ou des rangs) ; ce type de test n'exige aucune forme de distribution particulière et s'applique à de nombreuses situations, dont les échantillons de petite taille (inférieurs à 30 observations). Les plus utilisés sont les tests du *Chi-deux* et de *Kolmogorov-Smirnov*.

– Pour une variable non métrique qui correspond à un *choix exclusif* (réponse binaire telle que « oui / non », « possède / ne possède pas », etc.) dont la distribution statistique suit une loi binomiale qui peut être approximée par une loi normale, alors on peut utiliser un test paramétrique pour comparer des fréquences.
– Pour une variable métrique qui n'est *pas approximable par une loi normale*, on choisit un test non paramétrique adapté à la décision à prendre (le plus souvent : un test pour comparer l'écart entre médianes).

POUR ALLER PLUS LOIN – Présentation de quelques lois statistiques fondamentales

Loi normale ou **loi de Laplace-Gauss :** c'est la loi la plus connue et la plus utile parmi les lois de probabilités. Un grand nombre de variables (par exemple : la taille des femmes adultes d'une population donnée, la durée de vie des produits, le poids des graines de maïs, etc.) suivent une loi normale. Lorsqu'une variable donnée X suit une loi normale de paramètres μ et σ, on la note : X~N(μ;σ). La fonction « densité de probabilité » de la loi normale dépend de la moyenne de la distribution (μ) et de son écart type (σ) : elle se formule ainsi :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

La courbe de cette densité, symétrique par rapport à la moyenne (μ), varie de -∞ à +∞ et l'aire sous la courbe est égale à 1. Si une variable suit une loi normale, 68 % des valeurs sont comprises dans un intervalle autour de la moyenne de ± 1 σ, 95 % dans un intervalle de ± 2 σ et 99,7 % dans un intervalle de ± 3 σ autour de la moyenne.

Si la variable X suit une loi normale de moyenne $\bar{x}$ et d'écart-type σ, la variable centrée réduite $(\bar{x} - \mu)/\sigma$ suit aussi une loi normale, particulière et unique appelée loi normale centrée réduite de moyenne 0 et d'écart-type 1. Une seule table est alors suffisante pour calculer l'aire sous la courbe à gauche d'une valeur z ($P(Z \leq z)$). Quels que soient les paramètres μ et σ de la distribution normale de la variable X, on peut alors écrire que :

$$P(X \leq x) = P(Z \leq \frac{\overline{x} - \mu}{\sigma})$$

par conséquent, *P* peut être lu dans la table de la loi normale centrée réduite.

Loi binomiale : c'est la loi de probabilité d'une variable aléatoire discrète qui donne le nombre de succès lorsqu'on renouvelle n fois de manière indépendante une épreuve de Bernoulli de paramètre *p*. Lorsqu'une variable X donnée suit une loi binomiale de paramètres n et p, on la note : X~B(n;p). La probabilité individuelle (fonction de masse) est donnée par l'expression :

$$P(X = k) = \frac{n!}{(n-k)!k!} p^k (1-p)^{(n-k)}$$

☞

© Dunod – Toute reproduction non autorisée est un délit.

☞

Les probabilités individuelles $P(X = k)$ en fonction des paramètres n et p et les probabilités cumulées (fonction de répartition) $P(X \leq k)$ peuvent être lues dans les tables de la loi binomiale.

Il y a bien d'autres lois théoriques qui décrivent la distribution des variables aléatoires discrètes comme par exemple, la loi de Poisson, la loi uniforme, la loi géométrique, la loi hypergéométrique, la loi binomiale négative, etc.

Approximation (par une loi normale) : sous certaines conditions ($n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$), il est possible d'approcher la loi binomiale $X \sim B(n;p)$ par la loi normale $X \sim N\left(np;\sqrt{np(1-p)}\right)$. Étant donné qu'une loi de probabilité de variable discrète est approximée par une loi de probabilité de variable continue, il faudra prendre en compte la correction de continuité. Ainsi, $P(X = k) = P(k - 0{,}5 < X < k + 0{,}5)$, $P(X \leq k) = P(X \leq k + 0{,}5)$, etc.

1.2 Les échantillons de mesure à comparer

Le nombre d'échantillons à comparer – un seul, deux ou plusieurs simultanément – permet d'identifier le type de test (ajustement ou comparaison) et l'outil statistique adapté. De ce fait, on peut identifier trois classes de tests de comparaison : avec un seul, deux, ou plusieurs échantillons de mesures.

Un seul échantillon de mesure

Dans ce cas, la valeur observée sur l'échantillon est comparée à une valeur de référence (fréquence ou moyenne) connue : soit une norme théorique déterminée à partir d'une loi de distribution donnée ; soit une valeur de référence – un standard – fondée sur l'expérience, ou sur des spécifications techniques (comme en contrôle de qualité). Cette situation définit la classe des *tests de comparaison à une norme*.

Le *test d'ajustement* (appelé aussi *test d'adéquation)* relève de cette classe. Son but est de vérifier que la distribution statistique d'une variable est compatible avec une loi de distribution préétablie (dont la loi normale ou la loi binomiale).

Illustration pour un seul échantillon de mesure. Le cas « Cirage » comprend deux objectifs décisionnels qui reposent sur des informations recueillies auprès d'un seul échantillon. La **décision 1** s'appuie en apparence sur deux effectifs d'un seul échantillon qui correspondent au choix émis soit pour la boîte (90 répondants), soit pour le tube (110 répondants). Est-ce à dire que 110 étant supérieur à 90, le tube est préféré à la boîte en comparant directement 110 à 90 ? Non !

Les personnes sollicitées ont choisi entre l'un ou l'autre des deux conditionnements proposés. Il s'agit donc d'un *choix exclusif* (boîte ou tube) exprimé par un échantillon *unique*, lequel peut être représenté par la loi binomiale. Dans ce contexte, puisqu'il faut choisir entre le tube ou la boîte, si les deux conditionnements sont perçus comme équivalents, il y a donc une chance sur deux pour que l'un ou l'autre soit choisi. Cette probabilité de 0,50 correspond donc à l'absence de choix marqué qui traduit une indécision (il y a autant d'individus pour choisir chaque terme de l'alternative).

En fait, un choix se dessine dès lors qu'une des modalités de l'alternative s'écarte d'une chance sur deux d'être choisie : cette probabilité est la *norme* qui découle de la loi décrivant le phénomène étudié. L'objectif du test statistique utilisé est de montrer si l'écart constaté entre la norme et la valeur observée dans l'échantillon est bien significativement différent de 0,50.

En conséquence, la décision 1 du cas Cirage repose sur la seule comparaison des 110 choix exprimés pour le tube à la norme 100 (sur les 200 personnes testées, une moitié choisirait le tube et l'autre moitié la boîte en cas d'indécision) et non aux 90 choix exprimés pour la boîte. On peut tout aussi bien raisonner de manière réciproque en se demandant si les 90 choix exprimés pour la boîte sont significativement *inférieurs* à cette même norme. Ajoutons que, dans le cadre d'un choix exclusif, le nombre de modalités définissant ce choix peut, bien sûr, être supérieur à deux : s'il y a trois modalités, la norme devient donc une chance sur trois (0,333) et avec quatre modalités, une chance sur quatre (0,25), etc.

Pour décider s'il est opportun de substituer le tube à la boîte (objectif de la décision 1), il faut donc comparer, sur l'unique échantillon des 200 personnes testées, le nombre de choix en faveur du tube (110, soit 55 %) à celui traduisant l'absence de choix (100, soit 50 %).

De même, dans la **décision 2**, il n'y a qu'un seul échantillon de mesure (effectuée d'ailleurs sur le même échantillon d'individus que la décision 1). Cette mesure porte sur la fluidité perçue de la pâte à partir d'une échelle verbale en cinq modalités étalonnées (quasi échelle d'intervalle). On dispose donc de la distribution des réponses obtenues auprès des 200 personnes testées sur ces cinq valeurs (de 1 : « beaucoup trop fluide » à 5 « beaucoup trop dure ») composant la variable « Consistance » qui est métrique (quasi-échelle d'intervalle – *cf.* chapitre 1).

L'objectif du test est de vérifier que l'appréciation *moyenne* de la consistance de la pâte ne s'écarte pas de manière significative des attentes subjectives de l'échantillon d'individus étudiés et traduites par la modalité « juste bien » (valeur 3), laquelle prend alors valeur de référence. La décision 2 s'appuie donc, non sur une norme statistique, mais sur une « norme » empirique : la décision de modifier la texture de la pâte ou de la maintenir telle dépend de la comparaison de la moyenne calculée sur l'échantillon (2,45) à cette norme (3).

S'il n'y a pas de norme ou de référence préexistante, il n'y a pas de comparaison à envisager. Mais on peut chercher à estimer la valeur que, dans la population, la moyenne ou la fréquence (ou tout autre paramètre étudié) peut prendre : cela se fait à partir de sa valeur observée sur l'échantillon. Cette situation, très classique, correspond à l'*estimation par intervalle de confiance* : elle permet d'estimer pour tout paramètre statistique, la gamme plausible de ses valeurs inconnues sur la population à partir de ce qui est observé sur un échantillon.

Deux échantillons de mesure

Situation fréquente en gestion où deux fréquences ou deux moyennes sont comparées. Le test de comparaison diffère selon que les deux échantillons à comparer sont indépendants ou appariés.

© Dunod – Toute reproduction non autorisée est un délit.

Échantillons indépendants

Les tests de comparaison entre *échantillons indépendants* s'imposent si les deux groupes d'observations comparés sont exclusifs (on ne peut appartenir en même temps aux deux échantillons comparés) : la mesure est effectuée sur des groupes définis par des caractéristiques qui les différencient de manière exclusive (des hommes et des femmes ; des utilisateurs et des non-utilisateurs, etc.).

> **Illustration pour deux échantillons indépendants.** Le cas « Cirage » comprend deux objectifs décisionnels qui reposent sur des informations recueillies auprès de deux échantillons indépendants : la décision 3 conduit à comparer deux fréquences et la décision 4, deux moyennes.
> Bien que la **décision 3** soit formulée en des termes proches de la première décision (mais avec des caractéristiques sociales différentes), elle ne correspond pas à la même problématique décisionnelle : il s'agit de résoudre un problème de segmentation et non d'évaluer un niveau de préférence. En effet, l'objectif est de vérifier si l'accueil de la crème-cirage conditionnée en tube avec applicateur est meilleur chez une clientèle plus aisée [A] que modeste [M]. Comparer la fréquence du choix *pour le tube* exprimé par deux échantillons indépendants (un même individu ne peut appartenir à la fois à la classe « aisée » et à la classe « modeste ») permet de vérifier si une cible se montre plus intéressée qu'une autre. Il faut vérifier si l'appréciation pour le tube [T] diffère ou non dans les deux classes étudiées [A et M].
> Techniquement, il s'agit de comparer les 59 % des [A] [(130/220)*100] aux 55 % des [M] [(110/200)*100] ayant préféré le tube. Cet objectif implique donc de restructurer les données initiales pour construire un nouveau tableau qui intègre simultanément les données des situations 1 et 3. Ce serait une erreur de répliquer la solution statistique de la décision 1 pour comparer ses résultats directement avec ceux de la décision 3.
> L'objectif, dans la **décision 4**, est de s'assurer que la performance perçue du produit testé est indépendante du niveau d'utilisation du cirage. Il faut donc vérifier qu'il n'existe pas de différence d'appréciation entre les deux groupes étudiés, les petits (Pu) et les plus gros (Gu) utilisateurs.
> Techniquement, il s'agit de comparer les deux moyennes d'appréciation calculées sur deux groupes indépendants – respectivement 5,7 (Pu) et 6,3 (Gu) –, car les personnes testées ne peuvent être à la fois « petits » et « gros » utilisateurs : elles appartiennent exclusivement à une seule de ces deux classes. La décision 4 relève donc de la comparaison d'une moyenne entre deux échantillons indépendants (en prenant garde de choisir le bon outil statistique – test paramétrique ou non paramétrique – selon que les données respectent ou non certaines contraintes formelles).

Échantillons appariés

Les tests de comparaison sur échantillons appariés s'imposent si les deux groupes de mesure comparés ne sont pas indépendants : les valeurs obtenues sur le second échantillon de mesure dépendent des valeurs obtenues sur le premier.

L'appariement est direct, et parfait, s'il concerne les mêmes unités (personnes, objets…) qui subissent deux séries de mesures. Il est indirect si à chaque individu du premier échantillon, correspond dans le second échantillon un individu qui,

bien que différent, partage les mêmes caractéristiques de celui du premier échantillon.

> **Illustration pour deux échantillons appariés.** Dans le cas « Cirage », la **décision 5** relève de cette catégorie d'échantillons. En effet, l'objectif est de vérifier que le nouvel applicateur est perçu comme au moins équivalent, sinon supérieur, à celui utilisé habituellement par les enquêtés. Pour ce faire, le produit concurrent habituel est évalué lors de la première visite et le tube en test l'est aussi, après usage et auprès des mêmes personnes, lors de la seconde visite : les mesures sont appariées.
> Le produit testé étant évalué après usage par comparaison au produit usuel – lequel a valeur de référence –, sa notation n'est pas indépendante de l'appréciation du produit habituel. Les deux séries de mesure (successivement sur le produit usuel et sur le produit testé) sont donc liées.
> Techniquement, ce n'est donc pas la différence entre les deux moyennes qu'il faut analyser (cas des échantillons indépendants : 2,95 - 3,22), mais la moyenne de la différence entre l'évaluation mesurée pour le produit usuel d'abord, puis pour celui en test, sur l'ensemble de chacun des utilisateurs. C'est donc le test de comparaison d'une moyenne entre deux échantillons de mesures appariés qui est adapté pour cette situation, mais il implique de calculer, individu par individu, cet écart d'appréciation (ce que le tableau de synthèse fourni dans la situation 5 ne permet pas).

Plus de deux échantillons de mesure

Il est commun de devoir comparer plus de deux échantillons simultanément. Il y a deux manières de le faire.

La première, très analytique, consiste à comparer chaque paire d'échantillons de manière systématique avec les tests qui viennent d'être présentés et selon la même procédure. Il suffit ensuite de hiérarchiser leurs résultats en fonction de la significativité des tests pour tirer les conclusions opérationnelles qui s'imposent. La seconde manière, plus globale, consiste à vérifier si au moins un échantillon diffère significativement des autres et, dans l'affirmative, à n'opérer qu'ensuite les comparaisons par paire de manière systématique.

La comparaison de plusieurs échantillons correspond à une analyse bivariée, laquelle croise deux variables ayant chacune un statut propre. Si l'une est métrique et l'autre non, chacune des modalités de cette dernière variable définit en fait une classe de mesure, autrement dit, un échantillon. Il y a donc autant d'échantillons à comparer que de modalités de la variable non-métrique utilisée.

Les tests de comparaison multiple permettent de comparer les moyennes de plusieurs échantillons, avec des tests spécifiques selon que les échantillons sont indépendants ou appariés. Deux modèles statistiques s'imposent, selon la nature des variables étudiées : l'analyse de variance métrique ou test de Fisher-Snedecor et l'analyse de variance non métrique ou test de Kruskal-Wallis.

© Dunod – Toute reproduction non autorisée est un délit.

Illustration de comparaison multiple entre trois échantillons indépendants. La **décision 7** du cas « Cirage » correspond à cette situation : elle vise à choisir un argumentaire publicitaire convaincant parmi les trois proposés.
- Une variable, appelée « facteur », définit les échantillons : cette variable, le plus souvent non-métrique (ou « discrétisée » si elle est métrique d'origine), est utilisée pour caractériser les observations. Dans l'exemple, chacun des trois argumentaires constitue une modalité de la variable qui est donc, ici, nominale avec trois modalités.
- Une variable définit la mesure effectuée : c'est nécessairement une variable métrique distribuée normalement pour le test de Fisher-Snedecor ou une variable ordinale pour le test de Kruskal-Wallis. Dans l'exemple, la moyenne du critère d'« efficacité » est calculée par type d'argument : c'est bien une variable métrique

Compte tenu de tous ces éléments relatifs à la nature des variables et à celle des échantillons, le tableau 2.1 récapitule les tests de comparaison les plus usuels pour aider l'analyste à sélectionner dans le menu d'un logiciel de traitement statistique, le test le mieux adapté aux données disponibles et au type de décision à prendre.

2 Mettre en œuvre les procédures statistiques

Après avoir identifié le type de comparaison qui permet de répondre à la problématique décisionnelle à traiter, il convient de formuler les hypothèses statistiques à tester et de choisir, en conséquence, l'outil de calcul le plus adapté. En effet, l'une des difficultés rencontrée dans la réalisation d'un test de comparaison est de sélectionner dans le menu d'un logiciel de traitement, l'outil pertinent compte tenu de la nature des données.

2.1 Formuler l'hypothèse statistique à tester – Ho et H1 – en cohérence avec la décision managériale

Tous les tests de comparaison reposent sur une même logique formalisée dans le test d'hypothèse nulle, lequel consiste à vérifier l'absence d'écart entre les valeurs des paramètres statistiques étudiés (comme la fréquence ou la moyenne).

© Dunod – Toute reproduction non autorisée est un délit.

Tableau 2.1 – Les différents types de tests statistiques pour l'estimation et la comparaison

Type d'analyse	Échantillons d'observation		Nature des variables			Objectif statistique
			Non-métrique		Métrique	
	Nombre	*Type*	*Nominale*	*Ordinale*	*Intervalle ou de proportion*	
Univariée	*Un seul*		Chi-carré (X^2) Test binomial Test du signe	Chi-carré (X^2) Test de Kolmogorov-Smirnov	Intervalle de confiance Test z de la loi normale Test t de Student	***Estimation*** ***Ajustement***
Bivariée	*Deux*	Indépendants	Chi-carré (X^2) Test de Kolmogorov-Smirnov Test de Fisher	Chi-carré (X^2) Test de la médiane Test U de Mann-Whitney	Test z de la loi normale Test t de Student	***Comparaison***
		Appariés	Test de McNemar	Test du signe Test de Wilcoxon	Test t de Student (sur moyenne des écarts)	
	Trois et plus	Indépendants	Chi-carré généralisé Test de Kolgomorov-Smirnov	Test de la médiane Test H de Kruskal-Wallis	Analyse de la variance	
		Appariés	Test Q de Cochran	Test de Friedman	Analyse de la variance	

Par exemple, si deux échantillons quelconques sont bien issus d'une même population, il ne doit pas y avoir d'écart entre leurs deux moyennes comparées ; par hypothèse, cet écart est nul (d'où la notion d'hypothèse nulle). Mais, le plus souvent, en raison de l'erreur d'échantillonnage, cet écart n'est pas nul et diffère de zéro, aussi faible soit-il.

L'objectif du test de comparaison est d'apprécier l'amplitude de l'écart observé sur le paramètre étudié pour décider si cet écart résulte d'aléas (liés à l'échantillonnage) ou s'il est le signe d'une réelle différence entre les valeurs étudiées, indiquant que les échantillons ne proviennent pas de la même population de référence.

Si le test montre qu'il n'y a pas d'écart statistiquement significatif (hypothèse nulle vérifiée), on conclut que les deux échantillons appartiennent à la même population. Mais si l'écart est confirmé, comment l'interpréter ? À l'hypothèse nulle, il est donc nécessaire d'associer une hypothèse alternative qui définit la manière dont cet écart doit être interprété. Mais sa formulation précise dépend de la nature de la décision managériale à prendre.

Dans les conventions d'écriture, l'hypothèse nulle est symbolisée par « H_0 » et l'hypothèse alternative par « H_1 » (parfois « $H_{A\,»}$). L'hypothèse nulle est une hypothèse basée sur le principe de non-différence de sorte que les écarts observés puissent être attribués à la seule l'erreur d'échantillonnage. Ainsi, dans le cas de la comparaison de la moyenne inconnue d'une population (μ) à une valeur de référence (μ_0), H_0 postule que $\mu = \mu_0$; dans le cas d'un test de comparaison des moyennes inconnues de deux populations μ_1 et μ_2 : H_0 postule que $\mu_1 = \mu_2$, et ainsi de suite.

Pour comparer n'importe quel paramètre statistique appartenant à une population (moyenne, proportion, variance, etc., et symbolisé par θ : théta) avec une valeur de référence, le test d'hypothèse est formulé de la manière suivante :

$H_o : \theta_{pop} - \theta_{réf} = 0$

$H_1 : \theta_{pop} - \theta_{réf} \neq 0$

où $\theta_{réf}$ correspond à la valeur de référence et θ_{pop} à la valeur du paramètre inconnu de la population.

L'hypothèse alternative peut être affinée en posant : soit $\theta_{pop} - \theta_{réf} > 0$ ($\theta_{pop} > \theta_{réf}$), soit $\theta_{pop} - \theta_{réf} < 0$ ($\theta_{pop} < \theta_{réf}$). Elle peut donc être formulée de trois manières différentes : c'est la décision managériale à prendre qui va déterminer la formulation la plus adaptée. Le test est *bilatéral* lorsque le paramètre inconnu de la population ne doit pas être différent de la valeur de référence (d'où l'usage indifférent des signes « ≠ » ou « < > » pour formuler l'hypothèse alternative) ; ou *unilatéral* lorsque le paramètre inconnu de la population ne doit pas être, soit supérieure, soit inférieure, à la valeur de référence : l'hypothèse alternative s'écrit alors soit avec « > » (test unilatéral droit), soit avec « < » (test unilatéral gauche).

La formulation de l'hypothèse alternative H_1 dans le cas « Cirage »[1]

Dans la **décision 1**, la décision de substituer le tube à la boîte est possible si, dans la population des utilisateurs de cirage, la vraie proportion des choix exprimés pour le tube est significativement supérieure à la norme ($\pi_0 = 0{,}5$). Or, comme le décideur souhaite que le tube soit préféré à la boîte, il s'attend à ce que H_o soit rejetée au seul bénéfice de l'alternative où H_1 ; donc le test d'hypothèse sera ainsi formulé (test unilatéral droit) :

$H_0 : \pi = \pi_0$ (il faut comprendre comme $\pi \leq \pi_0$)

$H_1 : \pi > \pi_0$ ou $\pi - \pi_0 > 0$

Dans la **décision 2**, modifier la texture de la pâte ne s'impose que si elle est perçue vraiment comme trop fluide (> 3) *ou* trop dure (< 3). Pour n'avoir rien à modifier, le décideur s'attend à ce qu'elle soit perçue « juste bien » ($\mu_0 = 3$) et que H_0 soit validée (la moyenne inconnue de la population et la valeur de référence μ_0 sont égales). Aussi, très logiquement, l'hypothèse alternative H_1 est donc ici : $\mu \neq 3$. Il s'agit alors d'un test bilatéral. Le test d'hypothèse sera donc ainsi formulé :

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$ ou $\mu - \mu_0 \neq 0$ (comprendre $\mu < \mu_0$ ou $\mu > \mu_0$)

Dans la problématique de ciblage de la **décision 3**, le décideur envisage comme possible que son offre corresponde mieux au groupe le plus aisé [A] qu'à l'autre [M]. Il s'attend donc à ce que H_0 soit rejeté et, parmi les personnes qui choisissent le tube, que les 59,1 % de [A] (130/220) soient significativement supérieurs aux 55 % de [M] (110/200). D'où cette formulation (test unilatéral droit) :

$H_0 : \pi_A = \pi_M$ ou $\pi_A - \pi_M = 0$ (comprendre $\pi_A \leq \pi_M$)

$H_1 : \pi_A > \pi_M$ ou $\pi_A - \pi_M > 0$

Dans la **décision 4**, le décideur veut vérifier que l'appréciation du produit testé est identique quel que soit le niveau d'utilisation du cirage (donc vérifier H_o). Aussi, une quelconque différence significative en faveur des petits (Pu) ou des gros (Gu) utilisateurs l'obligerait à devoir admettre que l'appréciation varie selon le degré d'usage et à rejeter H_o (qu'il voudrait voir vérifiée). Ce test d'hypothèse (bilatéral) se formule donc comme suit :

$H_0 : \mu_{Pu} = \mu_{Gu}$ ou $\mu_{Pu} - \mu_{Gu} = 0$

$H_1 : \mu_{Pu} \neq \mu_{Gu}$ ou $\mu_{Pu} - \mu_{Gu} \neq 0$ (comprendre $\mu_{Pu} - \mu_{Gu} < 0$ ou $\mu_{Pu} - \mu_{Gu} > 0$)

Dans la **décision 5**, c'est H_1 qui permet d'évaluer la valeur concurrentielle de l'offre : il faut que l'appréciation du produit testé [T] ne soit pas inférieure à celle du produit usuel [U] (deux échantillons de mesure appariés obtenus auprès d'un

1. Cette partie expose les principes du raisonnement statistique liés à la formulation des hypothèses. Le traitement concret des données du cas « Cirage » est développé dans la section 2.

© Dunod – Toute reproduction non autorisée est un délit.

même échantillon d'individus, soit deux échantillons de mesure appariés). On teste donc la moyenne des différences de notation où l'on s'attend à ce que l'écart entre produits usuel et testé soit à la fois négatif (usuel < testé : test unilatéral gauche) et différent de 0 (rejet de H_0) :

$$H_0 : \mu_D = \mu_U - \mu_T = 0$$
$$H_1 : \mu_D < 0$$

Pour éviter toute erreur d'interprétation dans le résultat statistique du test sur des échantillons appariés, il importe d'être rigoureux dans le calcul de la différence en posant systématiquement : [valeur du premier échantillon de mesure – ici : le produit usuel] moins [valeur du second échantillon de mesure – ici : le produit testé].

2.2 Choisir l'outil de comparaison en fonction de la nature des données

La théorie statistique a permis de transformer l'écart entre la valeur observée et la valeur théorique en un paramètre qui exprime l'amplitude de la différence entre elles. C'est ce paramètre que les tests de comparaison analysent, mais il diffère selon l'outil utilisé (z, t de Student, F de Snedecor, Chi-carré, etc.) en fonction des conditions d'observation.

Chaque test repose sur l'une de ces statistiques ayant sa propre loi de distribution, et à partir de laquelle on peut calculer la probabilité d'observer cette valeur. Selon que cette probabilité est plus grande ou plus petite qu'une valeur seuil définie par l'analyste – appelée « seuil de risque » ou « seuil de signification » –, le statisticien accepte l'hypothèse nulle H_0 ou la rejette pour considérer l'hypothèse alternative H_1 comme plausible. Mais quel test utiliser selon le cadre de la décision à prendre ?

Le choix de l'outil statistique dépend, non pas de la décision, mais de la nature des variables (métriques ou non) et des conditions d'obtention des données traitées. La nature des échantillons et leur nombre définissent une méthode de comparaison (*cf.* § 1.1) ; mais c'est la nature paramétrique ou non paramétrique, et les caractéristiques (par ex. le type de distribution) des variables qui permettent de préciser l'outil adapté, soit paramétrique, soit non paramétrique. Le tableau 2.1 récapitule les différents tests de comparaison en fonction de la nature des variables, du nombre et de la nature des échantillons considérés.

Pour utiliser un test paramétrique, l'analyste doit identifier les paramètres de la population s'ils sont connus, ou les estimer à partir des paramètres de l'échantillon. L'objectif est de déterminer la loi de distribution de l'échantillonnage du paramètre testé. En fonction de cette loi de distribution, la valeur centrée réduite du paramètre suivra une loi normale centrée réduite, une loi de Student ou une loi quelconque.

Remarque

La *loi normale centrée réduite* est « centrée » car son origine est la moyenne (de valeur 0 donc) ; et « réduite » car l'écart-type est l'unité de mesure prise pour la mesurer. Toute variable de moyenne $\bar{x}$ et d'écart-type s peut être transformée en variable centrée réduite à l'aide de la formule $(x - \bar{x})/s$. Comment alors vérifier si la variable étudiée suit une loi normale ? D'une manière générale, on démontre qu'une variable métrique quelconque tend à suivre une loi normale si la taille de son échantillon est supérieure ou égale à 30 observations tirées selon les règles d'un tirage probabiliste, et ceci quelle que soit la distribution de la population (Théorème central limite).

Comment vérifier si une variable suit une loi normale

Quelle que soit la taille de l'échantillon, on peut le vérifier de manière pragmatique si :

– les paramètres de tendance centrale (moyenne, mode et médiane) ont des valeurs très proches (puisqu'elles se confondent si la variable suit une loi normale) ;
– la boîte à moustaches permet de vérifier graphiquement la symétrie de la distribution des données et la présence de valeurs extrêmes ;
– les observations sont alignées autour d'une droite dans le graphe quantile-quantile (Q-Q plot) ou sur la droite de Henry ;
– les valeurs standardisées des paramètres de forme (*cf.* ch. 1) – aplatissement (« Kurtosis ») et asymétrie (« *skewness* ») – sont elles-mêmes à l'intérieur d'un intervalle $\pm 2\sigma$;
– en utilisant un test adapté : parmi les tests statistiques qui vérifient la normalité d'une variable, celui de Shapiro-Wilk est le plus fiable, mais d'autres tests (Khi-carré, Kolmogorov-Smirnov et ses variantes) peuvent être utilisés.

Si les conditions de normalité sont remplies avec l'une ou l'autre de ces méthodes pragmatiques, on peut donc considérer que les variables étudiées se distribuent selon une loi normale et, en conséquence, utiliser un test paramétrique pour tester les écarts observés. Si ce n'est pas le cas, diverses transformations de la variable initiale (contenues dans les logiciels) peuvent être entreprises pour la rendre normale ; à défaut de transformation pertinente ou quand cette variable ne peut être approximée par la loi normale, l'usage d'un test non paramétrique s'impose.

Les transformations les plus usuelles

Trois types de transformations sont utilisés couramment : racine carrée, logarithme et arcsinus. La première est utilisée lorsqu'une variable X suit une loi de Poisson en calculant une nouvelle variable $X' = \sqrt{X + 0,5}$. La transformation logarithmique est utilisée quand une variable X contient beaucoup de valeurs trop faibles. La transformation est alors $X' = \log(X + 1)$. Quand les données brutes sont exprimées en fréquence (entre 0 et 1), on peut utiliser la transformation arcsinus de

© Dunod – Toute reproduction non autorisée est un délit.

la racine carrée : $X' = \arcsin(\sqrt{X})$. Il existe d'autres types de transformations (carré, exponentielle, inverse, etc.), mais leur développement dépasse le cadre de cet ouvrage.

Si la variable étudiée suit *une loi binomiale*, correspondant à un choix exclusif, sous certaines conditions peu contraignantes, elle peut être approximée par la loi normale (*cf.* § 1.1) et les écarts de fréquences testés avec un test paramétrique. Si cette approximation est impossible, un test non paramétrique s'impose.

Face à une *variable non métrique ordinale* qui génère des rangs ou des hiérarchies, il faut se demander si le caractère ordinal est un élément majeur dans la prise de décision. Dans l'affirmative, il faut donc choisir un test non paramétrique permettant bien de tester les éventuels écarts de rang observés.

2.3 Passer du résultat statistique à la prise de décision

Une fois choisi l'outil statistique qu'il convient d'appliquer, compte tenu du contexte décisionnel et des données traitées, il convient de différencier ce qui relève de la démarche statistique et ce qui relève de l'utilisation des résultats obtenus pour prendre la décision managériale ayant justifié l'usage d'un test de comparaison.

Parvenir à la décision statistique

La décision statistique – à bien différencier de la décision managériale – consiste à décider si l'hypothèse nulle H_0 est ou non acceptée au vu du résultat du test. Ce résultat est issu du calcul généralement effectué par le logiciel utilisé ; mais il dépend de choix préalables que l'analyste doit faire (d'ailleurs demandés par le logiciel) : l'importance de l'erreur qu'on accepte de commettre dans la décision statistique et le type de test choisi – unilatéral droit, unilatéral gauche ou bilatéral.

Définir le seuil de risque associé à la décision statistique

- Déclarer qu'un écart est « significatif » n'a de sens que par rapport à une référence. Cela signifie que cet écart ne peut relever du hasard s'il apparaît dans une proportion de cas supérieure à un pourcentage fixé *a priori*.
- Ce pourcentage est appelé « *seuil de risque* » ou *seuil de signification* (voire « seuil de probabilité »). Il correspond à la *probabilité de commettre une erreur en rejetant l'hypothèse nulle Ho* formulée (l'absence d'une différence significative entre moyennes comparées, par exemple) alors que dans la réalité cette hypothèse est fausse (car H_0 est vraie). Il est symbolisé par la lettre grecque α (alpha).
- Selon l'importance des enjeux liés à la décision, c'est l'analyste qui définit le seuil de risque : usuellement fixé à 5/100 (0,05), il peut être plus strict : 1/100 (0,01), voire 1/1000 (0,001), ou plus large : 10 % (0,10).

Remarque

L'usage consacré d'un seuil de risque de 5 % s'explique cependant. Le seuil de risque α correspond à l'erreur de première espèce. Une autre erreur, dite de seconde espèce et notée β (bêta), mesure la *probabilité de commettre une erreur en acceptant l'hypothèse nulle H_0 alors qu'elle est fausse*. Or, ce risque β est extrêmement élevé quand α est très petit, sauf à exploiter des échantillons de très grande taille, ce qui est rarement le cas. Le risque β reste raisonnable pour un seuil de risque α de .05 qui constitue un bon compromis pour la décision.

		La décision prise	
		H_0 retenue	H_0 rejetée
La réalité	H_0 vraie	Choix correct Seuil de confiance : $1 - \alpha$	Choix erroné Risque de 1re espèce : α
	H_0 fausse	Choix erroné Erreur de 2e espèce : β	Choix correct Puissance du test : $1 - \beta$

S'agissant d'une probabilité, le complément de ce seuil de risque est appelé *seuil de confiance* ou *niveau de confiance* : il correspond à la probabilité d'accepter l'hypothèse nulle alors qu'elle est vraie. Pour un seuil de signification (ou seuil de risque) de 5 %, le seuil de confiance est donc de 95 %. Seuil de confiance et seuil de risque délimitent respectivement les zones d'acceptation et de rejet de l'hypothèse nulle :

- Réduire le seuil de risque (en passant de 5 % à 1 %) *augmente* l'étendue de la zone d'acceptation, c'est-à-dire les limites de l'intervalle de confiance ;
- Augmenter le seuil de risque (en passant de 5 à 10 %) *réduit* l'étendue de la zone d'acceptation.

Un changement de seuil de risque peut donc conduire à modifier la décision concernant l'acceptation ou le rejet de l'hypothèse nulle. Le choix du seuil de risque n'est donc pas anodin et devrait toujours être fondé sur une réflexion liée à l'appréciation des coûts engendrés par une décision erronée suite aux observations effectuées sur l'échantillon étudié.

Exemple montrant l'importance du seuil de risque. Environ deux tiers (65 %) des clients « fidèles » d'une enseigne de distribution fréquentent aussi une autre enseigne concurrente. Dans le cadre d'une étude d'image concurrentielle, nous souhaitons travailler sur un échantillon respectant cette proportion. Or, dans l'échantillon des 440 individus tirés, 264 (60 %) fréquentent les deux enseignes. Peut-on néanmoins considérer que l'échantillon reste représentatif de la population de référence ?

Avec un seuil de risque de 1 % (soit un niveau de confiance de 99 %), l'étendue de la zone d'acceptation de H_0 serait d'après les calculs de 12 % [$P\,(0{,}54 < \pi < 0{,}66) = 0{,}99$)]. Comme la valeur de référence ($\pi_0 = 0{,}65$) est incluse dans cet intervalle ($0{,}65 \in [0{,}540\ ;\ 0{,}660]$), on en conclut que l'échantillon est bien représentatif de la population. Avec un seuil de risque de 5 % ($1 - \alpha = 0{,}95$), cette étendue est réduite à 9,2 % [$P\,(0{,}554 < \pi < 0{,}646) = 0{,}95$]. La valeur de référence étant extérieure à cet intervalle ($0{,}65 \notin [0{,}554\ ;\ 0{,}646]$), on devrait alors conclure que l'échantillon n'est pas représentatif de la population d'origine, mais correspond à une autre population de référence.

© Dunod – Toute reproduction non autorisée est un délit.

Décider de la nature du test : bilatéral ou unilatéral ?

Un *test bilatéral* (« two-tailed test ») permet de vérifier si la valeur étudiée se situe bien à l'intérieur des limites d'un intervalle centré autour d'une valeur centrale ayant valeur de référence. L'hypothèse H_o est privilégiée.

Exemple : calcul d'un test bilatéral. Revenons à l'exemple de la décision 2. La texture du cirage testé doit être juste équilibrée comme le souhaitent les clients : qu'il soit trop fluide (limite supérieure à ne pas franchir) ou pas assez fluide (limite inférieure à ne pas franchir), il ne sera pas satisfaisant. Un test bilatéral vise, ici, à vérifier si la vraie consistance testée se trouve bien à l'intérieur des limites qui peuvent être interprétées comme les attentes des prospects (sans se soucier si cette consistance est inférieure ou supérieure à celle souhaitée).

L'hypothèse alternative H_1 pour un test bilatéral est formulée de la manière suivante :

– pour une moyenne :

$$H_1 : \mu \neq \mu_0 \text{ ou } \mu - \mu_0 \neq 0 \text{ (comprendre } \mu < \mu_0 \text{ ou } \mu > \mu_0)$$

– pour une proportion :

$$H_1 : \pi \neq \pi_0 \text{ ou } \pi - \pi_0 \neq 0 \text{ (comprendre } \pi < \pi_0 \text{ ou } \pi > \pi_0)$$

Avec un *test unilatéral* (« one-tailed test »), le résultat est considéré différemment selon que la statistique étudiée est plus grande ou plus petite qu'une valeur de référence. Le décideur s'attache assez souvent à vérifier que la valeur observée dépasse cette valeur de référence ; l'hypothèse alternative H_1 est privilégiée.

Exemple de calcul d'un test unilatéral. Reprenons l'exemple des enseignes de la page précédente. Si l'on doit admettre que la proportion observée est plus petite que celle attendue, cela signifie que le nombre de clients « fidèles » aux deux enseignes considérées a baissé. Mais dans l'hypothèse où ce même pourcentage dépasse la valeur attendue, le nombre de clients communs aurait donc significativement augmenté.

Cette hypothèse alternative pour un test unilatéral est formulée de la manière suivante :

– pour une moyenne :

$$H_1 : \mu > \mu_0 \text{ ou } \mu - \mu_0 > 0 \text{ (test unilatéral droit)}$$

$$H_1 : \mu < \mu_0 \text{ ou } \mu - \mu_0 < 0 \text{ (test unilatéral gauche)}$$

– pour une proportion :

$$H_1 : \pi > \pi_0 \text{ ou } \pi - \pi_0 > 0 \text{ (test unilatéral droit)}$$

$$H_1 : \pi < \pi_0 \text{ ou } \pi - \pi_0 < 0 \text{ (test unilatéral gauche)}$$

Alors que le statisticien cherche à valider l'hypothèse nulle, le décideur peut être dans la situation inverse où c'est le rejet de H_0 qui l'intéresse, au bénéfice d'une hypothèse alternative unilatérale qui correspond à sa problématique et qu'il faut donc formuler sans ambiguïté. Mais cela ne peut se faire qu'en fixant clairement des règles de décision.

Les règles pour formuler la décision statistique

La difficulté à prendre la bonne décision managériale avec l'usage d'un test de comparaison découle de deux difficultés : l'explicitation initiale du problème managérial qui oriente le type de test à utiliser (comme vu au § 1.1) et la bonne interprétation des résultats du test appliqué, plus précisément les conditions d'acceptation ou de rejet de l'hypothèse nulle. Pour ce faire, l'analyste dispose de plusieurs critères complémentaires – mais bien différenciés pour interpréter le résultat du test statistique : l'intervalle de confiance, les valeurs critiques, les rapports critiques et la probabilité critique.

Décider sur la base de la valeur du rapport critique (souvent abrégé en « valeur critique »)

On **accepte** l'hypothèse nulle H_0 si la valeur critique calculée du paramètre du test est inférieure à la valeur critique déterminée dans la table du test pour une probabilité critique α (alpha) choisie par l'analyste (les tables proposent le plus souvent trois niveaux de probabilité α : 1 %, 5 %, 10 % ; soit en probabilité : 0,01, 0,05 et 0,10). Concrètement, pour un seuil α déterminé, si la valeur du test calculée varie entre 0 et la valeur limite fournie par la table, on accepte H_0. Cela signifie que l'écart est faible et n'est pas significatif d'une différence, compte tenu des conditions d'observation : puisque la différence entre les moyennes ou les fréquences testées n'est pas statistiquement significative, *les deux échantillons appartiennent donc à la même population de référence* et ne peuvent être différenciés.

On **rejette** l'hypothèse nulle H_0 pour accepter l'hypothèse alternative H_1 comme étant plausible dans le cas inverse : la valeur critique calculée est supérieure à la valeur critique lue dans la table du test au seuil de risque α (alpha) fixé. Concrètement, pour un seuil déterminé, si la valeur du test calculée est supérieure à la valeur limite fournie par la table, on rejette H_0 au bénéfice de H_1. Compte tenu des conditions d'observation, l'écart constaté est trop élevé pour ne pas être significatif d'une différence. Cette différence entre les moyennes ou les fréquences étant statistiquement significative, *les deux échantillons n'appartiennent donc pas à la même population de référence* et ne peuvent pas être confondus.

Décider sur la base de la probabilité critique

On **accepte** l'hypothèse nulle H_0 si la probabilité critique calculée pour le paramètre du test est supérieure au seuil de risque α (alpha) arrêté par l'analyste. Pour un seuil fixé à 5 %, si la probabilité calculée par le test varie entre 5 % et 100 % (0,05 et 1 en probabilité), on accepte H_0. Cela signifie que la probabilité d'observer l'écart constaté est élevée, compte tenu des conditions d'observation. La différence entre les moyennes ou les fréquences *n'étant pas statistiquement significative, les deux échantillons appartiennent à la même population de référence* et ne peuvent donc pas être différenciés.

© Dunod – Toute reproduction non autorisée est un délit.

On **rejette** l'hypothèse nulle H_0 pour accepter l'hypothèse alternative H_1 dans le cas inverse, quand la probabilité critique calculée est inférieure au seuil de risque fixé. Concrètement, pour un seuil de 5 %, si la probabilité calculée par le test varie entre < 5 % et 0 % (0,05 et 0,00 en probabilité), on rejette H_0 au bénéfice de H_1. Compte tenu des conditions d'observation, la probabilité d'observer l'écart constaté est très faible : la différence constatée entre les moyennes ou les fréquences est statistiquement significative ; comme *les deux échantillons n'appartiennent pas à la même population de référence*, ils ne peuvent pas être confondus.

Tableau 2.2 – Logique des décisions dans un test hypothèse

Point de référence	Conditions pour accepter H_0	Conditions pour rejeter H_0
Valeur critique (V_{Crit})	V crit.calculée < V crit.théorique	V crit.calculée > V crit.théorique
Probabilité critique (P_{Crit})	P critique calculée > α	P critique calculée < α

Pour certaines statistiques de comparaison (t, F, x^2), *la probabilité critique est calculée en fonction du nombre de degrés de liberté (abrégé en «* d.d.l. *» ou «* ν *» :* nu) lié au nombre d'observations disponibles. Ces degrés de liberté correspondent au nombre effectif d'observations nécessaires pour reconstituer toute une série de données dont le total est connu. Les « d.d.l. » définissent donc le nombre de comparaisons indépendantes fournies par les données traitées.

Décider sur la base d'une analyse de l'intervalle de confiance

Si les variables utilisées pour la comparaison sont approximables par une loi normale, l'intervalle de confiance peut être utilisé de manière pragmatique pour prendre la décision statistique.

- Dans le *cas de la comparaison à une norme*, on accepte l'hypothèse nulle si le paramètre étudié (moyenne ou fréquence…) appartient à l'intervalle de confiance de cette norme : l'échantillon appartient bien à la population de référence. Dans le cas où la valeur du paramètre étudié est en dehors de l'intervalle de confiance de la norme, on rejette Ho au bénéfice de H_1 pour conclure que l'échantillon utilisé appartient à une autre population que celle de la norme.
- Dans le *cas de deux échantillons indépendants*. Si les intervalles de confiance se chevauchent ou partagent une zone commune, on en déduit que les échantillons appartiennent à la même population de référence : Ho est donc acceptée. Dans le cas où les deux intervalles ne se recoupent pas, on rejette Ho au bénéfice de H_1 : les deux échantillons n'appartenant pas à la même population de référence, il importe donc de les différencier.
- Dans le *cas de deux échantillons appariés*, le raisonnement porte sur la moyenne des différences entre les paires de mesures. Cette moyenne est donc nulle s'il n'y a aucun écart au sein de chaque paire. Si l'intervalle de confiance de cette

moyenne contient la valeur 0 (zéro), on accepte l'hypothèse nulle Ho. On la rejette au bénéfice de H_1 si cette valeur 0 est extérieure à cet intervalle pour conclure que la différence entre les deux séries de mesure est bien significative.

Section 2 PASSER DU RÉSULTAT STATISTIQUE À LA DÉCISION MANAGÉRIALE : TRAITEMENT DU CAS « CIRAGE »

Après l'exposé, dans la section 1, des éléments qui prévalent dans la mise en œuvre d'un test de comparaison, nous organisons cette seconde section en présentant, pour chaque décision du cas, la justification des tests utilisés, les hypothèses statistiques et le calcul des paramètres liés aux tests, puis la manière d'interpréter les résultats, avant de passer de la décision statistique (acceptation ou rejet de l'hypothèse nulle) à la recommandation managériale[1].

1 La comparaison à une norme : un seul échantillon – le cas d'une proportion (décision 1)

1.1 Problématique décisionnelle et type de test statistique

Le risque commercial serait de lancer le tube avec son nouvel applicateur alors que la boîte lui serait préférée. Il faut donc commencer par vérifier l'attrait de ce tube : c'est l'objet du test d'accueil où, après usage, les utilisateurs actuels ont indiqué leur préférence soit pour la boîte, soit pour le tube de *manière exclusive*.

Deux offres *exclusives* étant mises en concurrence, il faut vérifier si la proportion observée des utilisateurs qui ont opté pour le tube (110/200 = 0,55) est statistiquement différente d'une chance sur deux dans la population (100/200, soit 50 %). Cela revient à *comparer la fréquence du choix émis à une référence* qui prend valeur de norme (ici la loi binomiale, puisque le choix est exclusif).

L'erreur serait de comparer le choix du tube directement à celui de la boîte (55 % à 45 %) : ces deux choix sont redondants puisque complémentaires. Le choix pour le tube doit être comparé à l'absence de choix (probabilité égale de choisir entre le tube ou la boîte).

1. Les sept situations décisionnelles du cas « Cirage » sont exposées dans le cas d'illustration qui ouvre le chapitre ; l'analyse du lien entre contexte décisionnel, implications méthodologiques et les échantillons est traité en section 1.

© Dunod – Toute reproduction non autorisée est un délit.

1.2 Formulation des hypothèses et résultats des tests applicables

Données	
Effectif de n = 200 ; proportion observée pour le *Tube* : $p = 0{,}55$; norme théorique (norme ou standard) $\pi_0 = 0{,}50$	
Hypothèse nulle et hypothèse alternative (test unilatéral droit) :	
$H_0 : \pi = \pi_0$ (il faut lire ($\pi \leq \pi_0$) $H_1 : \pi > \pi_0$	
Résolution selon le type de test utilisé :	
Formule :	Calcul et résultats au seuil α : *0,05*
avec la valeur critique	
$VC = \pi_0 + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$	$VC = 0{,}5 + 1{,}64\sqrt{\frac{0{,}55(1-0{,}55)}{200}} = 0{,}558$
avec le test de comparaison (rapport critique)	
$z = \frac{p-\pi_0}{\sqrt{\frac{p(1-p)}{n}}}$	$z = \frac{0{,}55-0{,}50}{\sqrt{\frac{0{,}55(1-0{,}55)}{200}}} = 1{,}421$
avec un test non paramétrique : le test d'ajustement	
$\chi^2 = \sum \frac{(np-nf)^2}{nf}$	$\chi^2 = \frac{(110-100)^2}{100} + \frac{(90-100)^2}{100} = 2$

On applique le test de comparaison à une norme théorique fondé sur les seuls résultats obtenus pour le tube [T] où un test unilatéral droit s'impose, puisqu'une préférence marquée en faveur du tube est espérée. L'utilisation d'un test paramétrique est possible puisque les trois conditions sont remplies ($n = 200 \geq 30$, $np = 110 \geq 5$ et $np(1 - p) = 49{,}5 \geq 5$). La probabilité critique associée à cette valeur du test z est de 0,0776 (7,76 %) ; elle est donc supérieure au seuil de risque de 5 % qui a été fixé.

1.3 Formulation des conclusions et recommandations

Selon le test utilisé, la conclusion reste la même :

1. La valeur de la norme ($\pi_0 = 0{,}50$) étant incluse dans l'intervalle de confiance de la fréquence observée inconnue de la population, ce résultat conduit à accepter l'hypothèse nulle.
2. Dans le test en z, la valeur de la probabilité critique étant supérieure au seuil de risque fixé (0,0776 > 0,050), l'hypothèse nulle ne peut être rejetée pour un seuil de confiance de 95 % et on peut conclure que les données de l'échantillon sont compatibles avec l'hypothèse H_0 (où $\pi = 0{,}50$). Par ailleurs, la valeur critique calculée de z (1,421) est inférieure à la valeur critique théorique que fournit la table de la loi normale : 1,64 au seuil de risque de 5 %.

3. Enfin, le test d'ajustement donne une valeur de x^2 inférieure à la valeur seuil pour un seul degré de liberté ($x^2_{\text{calculé}} = 2 < x^2_{\text{théorique}} = 3{,}84$).

Conclusion : *La proportion observée sur l'échantillon ne diffère donc pas significativement de la proportion théorique correspondant à une chance sur 2 de choisir le tube* (un seul de ces tests aurait suffi pour décider).

Qu'il y ait plus de personnes pour préférer le tube ne correspond donc pas à une réelle préférence et l'écart observé à la norme (une chance sur deux) n'est lié qu'aux aléas d'échantillonnage. Le décideur a donc toute latitude pour choisir en toute responsabilité entre la boîte B ou le tube T, selon des critères autres que statistiques. Et, comme les résultats ne permettent pas de conclure à une supériorité (ni à une infériorité) du tube, rien ne s'oppose donc à envisager le lancement du tube sur la base de cet écart de préférence, laquelle ne pourra cependant assurer à elle seule le succès commercial du produit.

Remarques

La résolution de ce type de problème n'implique pas de disposer des données individuelles ni d'effectuer les calculs. Certains logiciels de traitement statistique (tel Stagraphics) proposent un sous-programme nommé « tests d'hypothèses » qui permet d'obtenir directement un résultat. En indiquant l'hypothèse nulle, le type d'hypothèse alternative, la taille de l'échantillon et le seuil de risque choisi, on dispose directement de la probabilité critique liée à ces données pour conclure sur le rejet ou non de l'hypothèse nulle. Concernant l'incidence du choix du seuil de risque sur la décision : pour un seuil de risque usuel de 5 % (0,05), on note que la proportion observée (p = 0,55) est inférieure à la valeur critique du test (VC = 0,558) ; le choix du tube ne diffère pas d'un choix aléatoire (une chance sur deux ici).

2 La comparaison à une norme : un seul échantillon – le cas d'une moyenne (décision 2)

2.1 Problématique décisionnelle et type de test statistique

Avant de commercialiser le tube, le décideur veut s'assurer que la consistance perçue de la pâte satisfait la cible. L'échelle verbale quasi-métrique utilisée permet de mesurer l'adéquation du produit aux attentes individuelles. L'idéal serait que tous les répondants estiment cette consistance comme « juste bien » (note 3). L'objectif est donc de vérifier que la moyenne de l'appréciation ne s'écarte pas trop de cette référence cotée 3, car ce « juste bien » traduit les attentes des utilisateurs.

Sur le plan statistique, il s'agit de vérifier si la consistance de la pâte perçue par la population des utilisateurs (la vraie moyenne inconnue μ) est égale à la moyenne attendue ($\mu_0 = 3$). Cette situation correspond au *test bilatéral de comparaison d'une moyenne de population* (estimée par l'échantillon : ici 2,45) à *une valeur de référence* (ici : 3).

© Dunod – Toute reproduction non autorisée est un délit.

2.2 Formulation des hypothèses et résultats des tests applicables

Données	
Effectif : $n = 200$; norme théorique : $\mu_0 = 3$; moyenne observée : $\bar{x} = 2{,}45$; médiane : *Mé* = 2,5 ; écart – type : $s = 0{,}867$. Il s'agit ici d'un test bilatéral.	
Hypothèse nulle et hypothèse alternative	
$H_0 : \mu = \mu_0$ ou $\mu - \mu_0 = 0$ $H_1 : \mu \neq \mu_0$ (comprendre $\mu<\mu_0$ ou $\mu>\mu_0$)	
Résolution selon le type de test utilisé :	
Formule :	Calcul et résultats au seuil α : 0,05
Intervalle de confiance au seuil α : 0.05	
$IC_\mu : \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n-1}}$	$[2{,}45 - 1{,}96 \frac{0{,}867}{\sqrt{200-1}} ; 2{,}45 + 1{,}96 \frac{0{,}867}{\sqrt{200-1}}]$ $[2{,}330;2{,}570]$
Test de comparaison par le rapport critique	
$z_{\bar{x}} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n-1}}}$	$z_{\bar{x}} = \frac{2{,}45 - 3}{\frac{0{,}867}{\sqrt{200-1}}} = -8{,}949$ probabilité critique associée : $1{,}795 \times 10^{-19} << 0{,}05$
Test non paramétrique : test d'ajustement	
$\chi^2 = \Sigma \frac{(np - nf)^2}{nf}$ avec $v = k - 1$	$\chi^2 = 60{,}434 > 9{,}49$ pour $v = 4$ $\frac{(30-13{,}4)^2}{13{,}4} + \frac{(0-13{,}4)^2}{13{,}4} + \frac{(70-48{,}4)^2}{48{,}4} + \frac{(20-48{,}4)^2}{48{,}4} + \frac{(80-76{,}4)^2}{76{,}4}$

Remarques

L'estimateur sans biais de l'écart-type inconnu de la population est $s^* = \sqrt{n/(n-1)}\ s$: cela permet de calculer l'écart-type (ou « erreur-standard ») de la distribution d'échantillonnage ($\sigma_{\bar{x}}$) avec la formule $\sigma_{\bar{x}} = s/\sqrt{n-1} = 0{,}867/\sqrt{200-1} = 0{,}0615$. De plus, la taille de l'échantillon étant supérieure à 30, alors, selon le théorème central limite, la distribution d'échantillonnage de la moyenne suit approximativement une loi normale.

Que la pâte soit perçue trop dure ou trop fluide, ce n'est pas satisfaisant : la décision implique donc un test de comparaison *bilatéral*. Les caractéristiques de l'échantillon et les paramètres de forme (asymétrie et aplatissement) restant dans les normes d'une distribution normale, il est possible d'utiliser des tests paramétriques. Mais la valeur des moyenne et médiane observées (2,45) donnent à penser que cette distribution s'écarte de la norme fixée (3). Un test d'ajustement (non paramétrique) permet de vérifier si la distribution observée est ou non compatible avec celle de la norme pour les 5 modalités de réponses de l'échelle.

1. Les limites de l'intervalle de confiance de la moyenne observée sont comprises entre 2,33 et 2,57 : au seuil de confiance de 95 %, la norme (3) n'est pas incluse dans cet intervalle. Cela conduit à rejeter l'hypothèse nulle.

2. La statistique du test (rapport critique) $z_{\bar{x}} = (\bar{x} - \mu)/\ \sigma_{\bar{x}} = (2{,}45 - 3)/0{,}0615 = -8{,}949$ suit une loi normale centrée réduite . Dans l'absolu, sa valeur est très supérieure à la valeur théorique attendue : 1,96 au seuil de risque α de 5 % (lue dans la table de la loi normale). L'hypothèse nulle doit être rejetée.
3. Enfin, la probabilité critique α associée à cette valeur du test : $1{,}795 \times 10^{-19}$ est très faible, très inférieure donc au seuil de risque α de 5 % fixé. L'hypothèse nulle doit être rejetée au seuil de confiance de 95 %.

2.3 Formulation des conclusions et recommandations

Quel que soit le test utilisé (et un seul suffit), la conclusion reste la même et conduit à rejeter l'hypothèse nulle au bénéfice de l'hypothèse alternative. Et le recours à un test d'ajustement conduit au même résultat.

Le test d'ajustement compare la distribution théorique des observations pour une variable continue discrétisée en cinq classes, à la distribution observée[1]. Pour $k - 1$ degré de libertés (5 -1) où k correspond au nombre de modalités de réponses, la valeur critique théorique du χ^2 vaut 9,49, très nettement inférieure à celle calculée (60,43). La distribution des observations de l'échantillon n'est pas compatible avec la distribution des observations théoriques de la norme.

Conclusion. Les données de l'échantillon ne sont pas compatibles avec l'hypothèse H_0 où $\mu = 3$: la moyenne observée sur l'échantillon diffère très significativement de la norme attendue. Le signe négatif associé à la valeur critique (-8,949) indique bien que la pâte du cirage est trop fluide (orientation vers les faibles valeurs de la variable, c'est-à-dire « trop fluide ») : il faut donc en modifier la consistance en la durcissant (mais ce test n'indique pas dans quelle proportion pour obtenir une densité adaptée aux attentes…).

Remarque

À l'instar du cas d'une fréquence illustré en 1.3, certains logiciels de traitement permettent de tester directement une moyenne en fournissant les informations nécessaires (norme ; moyenne, écart-type et taille de l'échantillon ; seuil de risque α).

1. Pour une distribution normale, selon le nombre de classes, les proportions par classes sont les suivantes, pour 5 classes : 6.7%, 24.2%, 38.2%, 24.2%, 6.7% ; pour 7 classes : 4.8%, 11.1%, 21.2%, 25.8%, 21.2%, 11.1%, 4.8% ; pour 9 classes : 4.0%, 6.6%, 12.1%, 17.5%, 19.6%, 17.5%, 12.1%, 6.6%, 4.0%. Toutes ces valeurs sont déduites de la loi normale ; l'application de ces proportions à la taille d'un échantillon définit l'effectif théorique de chacune des classes d'une distribution normale.

© Dunod – Toute reproduction non autorisée est un délit.

3 La comparaison entre deux échantillons indépendants – le cas d'une fréquence (décision 3)

3.1 Problématique décisionnelle et type de test statistique

L'objectif est de vérifier si l'accueil du tube avec applicateur est meilleur dans une clientèle aisée [A] que dans une clientèle plus modeste [M]. Bien que cette situation évoque la première décrite, elle ne correspond pas au même objectif : il s'agit ici de traiter un problème de segmentation (et non de préférence). Il faut donc vérifier si le choix pour le tube diffère selon les deux classes étudiées : pour [A], on utilise les données de la décision 1 et pour [M], celles de la décision 3.

Techniquement, il s'agit donc de comparer les 59,1 % des [A] [(130/220)*100] aux 55 % des [M] [(110/200)*100] ayant préféré le tube. La solution consiste à comparer sur une *même* variable (la préférence) les résultats fournis par les deux groupes [A et M] indépendants (les enquêtés ne peuvent appartenir à la fois aux deux). La *comparaison des fréquences de deux échantillons indépendants* permet de résoudre le problème posé.

3.2 Formulation des hypothèses et résultats des tests applicables

Données

Effectif $n_A = 220$; fréquence du choix pour le Tube $p_A = 0,59$
Effectif $n_M = 200$; fréquence du choix pour le Tube $p_M = 0,55$

Hypothèses nulle et alternative *(test unilatéral droit)*

$H_o : \pi_A - \pi_M = 0$ (comprendre comme $d = \pi_A - \pi_M \leq 0$)
$H_1 : \pi_A - \pi_M > 0$

Test de comparaison *au seuil* $\alpha = 0,05$

On note : $d = p_A - p_M$

$$\sigma_d = \sqrt{p(1-p)(\frac{1}{n_A} + \frac{1}{n_M})}$$

avec

$$p = \frac{n_A p_A + n_M p_M}{n_A + n_M}$$

$$z = \frac{p_A - p_M}{\sigma_d} \sim N(0,1)$$

détail des calculs :

$$p = \frac{220x0,59 + 200x0,55}{220 + 200} = 0,571$$

$$\sigma_d = \sqrt{0,571(1-0,571)(\frac{1}{220} + \frac{1}{200})} = 0,0484$$

$$z_d = \frac{0,59 - 0,55}{0,0484} = 0,827$$

Probabilité critique associée :
$p - value = p(z_d > 0,827) = 0,204$

Donc : $p - value = 0,204 > 0,05$

Les conditions d'application pour effectuer un test paramétrique sont réunies : $n_A p$, $n_A (1 - p)$, $n_M p$, $n_M (1 - p)$ sont tous supérieurs à 5. Pour réserver le tube au seul segment des consommateurs les plus aisés, il faut que le choix du tube par la classe

[A] soit significativement supérieur à celui de la classe [M] ; un test *unilatéral droit* s'impose et le commanditaire s'attend à pouvoir rejeter H_0 au bénéfice de H_1.

3.3 Formulation des conclusions et recommandations

Au seuil de risque de 5 %, la valeur absolue de z (0,827) est très inférieure à la valeur seuil (1,645) et la probabilité critique qui lui est associée (0,204 soit 20,4 %) est très élevée. En conséquence, sur la base de l'échantillon étudié, l'hypothèse nulle ne peut être rejetée : la vraie différence de choix entre les deux populations n'étant pas significativement différente de 0, la structure des préférences est donc la même dans les deux populations.

Conclusion. L'objectif était de valider une éventuelle segmentation des clientèles selon un critère socio-économique. Les résultats du test montrent que le choix ne diffère pas selon la classe. Il n'est donc pas justifié de conserver la boîte pour les plus modestes [M] et de réserver le tube aux plus aisés [A].

Remarque

Comme pour la comparaison à une norme, la résolution de ce problème n'implique pas de disposer des données individuelles ni d'effectuer manuellement les calculs. Un sous-programme de « tests d'hypothèses pour la comparaison » proposé par certains logiciels permet d'obtenir directement le résultat. En indiquant la valeur de l'hypothèse nulle (0) et pour chaque échantillon : la fréquence observée et la taille de l'échantillon, puis le type d'hypothèse alternative et le seuil de risque choisi, on dispose directement des résultats pour conclure sur le rejet ou non de l'hypothèse nulle.

4 La comparaison entre deux échantillons indépendants – le cas d'une moyenne (décision 4)

4.1 Problématique décisionnelle et type de test statistique

Pour vérifier que la performance perçue du nouvel applicateur ne dépend pas du niveau d'utilisation du cirage, comme le pense le chef de produit, il faut comparer sur les deux échantillons indépendants que sont les petits (*Pu*) et les plus gros (*Gu*) utilisateurs, la moyenne du score composite créé. La problématique statistique est donc comparable à la précédente, mais appliquée sur une moyenne.

Comme le décideur s'attend à ce qu'il n'y ait pas de différence entre les deux moyennes observées – 5,7 (*Pu*) et 6,3 (*Gu*) –, techniquement il s'agit de vérifier que l'écart observé (0,6) n'est pas statistiquement significatif. La *comparaison d'une moyenne entre deux échantillons indépendants* répond à cet objectif, mais avec des tests différents selon que certaines contraintes théoriques sont ou non satisfaites.

© Dunod – Toute reproduction non autorisée est un délit.

4.2 Mise en œuvre de la procédure statistique

Données

Effectif $n_{Pu} = 160$; moyenne : $\bar{x}_{Pu} = 5,70$; écart-type : $s_{Pu} = 1,103$; médiane : *Mé* = 5,5 ; aplatissement standardisé : – 0,582 ; asymétrie standardisée : 3,196 ;
Effectif $n_{Gu} = 260$; moyenne : $\bar{x}_{Gu} = 6,307$; écart-type : $s_{Gu} = 1,2$; médiane : *Mé* = 6,0 ; aplatissement standardisé : – 3,636 ; asymétrie standardisée : 1,831 ; Intervalle de confiance de la moyenne : *Pu* [5,529 ; 5,871] ; de la moyenne : *Gu* [6,161 ; 6,453].

Hypothèses nulle et alternative *(test bilatéral)*

H_0: $\mu_{Pu} - \mu_{Gu} = 0$
H_1: $\mu_{Pu} - \mu_{Gu} \neq 0$ (comprendre $\mu_{Pu} - \mu_{Gu} < 0$ ou $\mu_{Pu} - \mu_{Gu} > 0$)

Intervalle de confiance de la différence au seuil d'erreur $\alpha = 0,05$

Notons : $d = \bar{x}_{Pu} - \bar{x}_{Gu}$

$$\sigma_d = \sqrt{\frac{s_{Pu}^2}{n_{Pu}} + \frac{\hat{s}_{Gu}^2}{n_{Gu}}}$$

$$IC_{\mu_{Pu}-\mu_{Gu}} : (\bar{x}_{Pu} - \bar{x}_{Gu}) \pm z_{\alpha/2}\sigma_d$$

Détail des calculs :

$$\sigma_d = \sqrt{\frac{1,103^2}{160} + \frac{1,2^2}{260}} = 0,1146$$

$$IC_{\mu_{Pu}-\mu_{Gu}} : (5,7 - 6,307) \pm 1,96 * 0,1146$$

$$IC_{\mu_{Pu}-\mu_{Gu}} : -0,607 \pm 1,96 * 0,1146$$

$$IC_{\mu_{Pu}-\mu_{Gu}} : [-0,832; -0,382]$$

Test de comparaison

$$z_d = \frac{\bar{x}_{Pu} - \bar{x}_{Gu}}{\sqrt{\frac{s_{Pu}^2}{n_{Pu}} + \frac{s_{Gu}^2}{n_{Gu}}}} \sim N(0,1)$$

$$z_d = \frac{5,7 - 6,307}{0,1146} = -5,295$$

probabilité critique associée : $5,956 \times 10^{-8} < 0,05$

Tests non paramétriques fondés sur la médiane :
test de Mann-Whitney *ou* test de Kolmogorov-Smirnov

- Les limites de l'intervalle de confiance de la différence entre les deux moyennes d'appréciation [0,06] sont comprises entre -0,832 et -0,382 ; au seuil de confiance de 95 %, cet intervalle ne contient pas la différence théorique [0] qui correspond à l'hypothèse nulle.
- La valeur absolue du test z (-5,295) est supérieure à la valeur test théorique (1,96).
- La probabilité critique α associée à cette valeur du test ($5,956 \times 10^{-8}$) est très inférieure au seuil de risque fixé de 5 %.

4.3 Formulation des conclusions et recommandations

Selon le test utilisé (intervalle de confiance de la différence ou test z), la conclusion reste la même : au seuil de risque de 5 %, tous les résultats conduisent à rejeter l'hypothèse nulle au bénéfice de l'hypothèse alternative. La moyenne des échantillons est donc significativement différente de 0 : ils n'appartiennent pas à une même population.

Ces résultats permettent de conclure que, contrairement à l'intuition du chef de produit, la performance de l'applicateur n'est pas perçue de manière identique selon le degré d'utilisation du cirage : les gros utilisateurs en ont une meilleure appréciation ($\bar{x}_{Gu} = 6{,}307$) que les petits utilisateurs ($x_{Pu} = 5{,}7$). Il devient alors envisageable de s'appuyer sur le segment des gros utilisateurs pour assurer le développement commercial et la promotion de la nouvelle formule.

Remarques

a. Concernant le raisonnement construit de manière pragmatique sur l'*intervalle de confiance* : pour un seuil de risque usuel α de 5 %, l'intervalle de confiance pour la moyenne de chaque échantillon est respectivement :

– échantillon P ($\bar{x}_{Pu} = 5{,}7$) : $5{,}529 < \mu_{Pu} < 5{,}871$

– échantillon G ($\bar{x}_{Gu} = 6{,}307$) : $6{,}161 < \mu_{Gu} < 6{,}453$

Or, on note que la borne *inférieure* de l'échantillon *Gu* (6,161) et la borne *supérieure* de d'échantillon *Pu* (5,871) ne se chevauchent pas et sont mutuellement exclues : il y a donc une différence entre les deux fréquences ; l'appréciation de la performance du tube est meilleure chez les gros utilisateurs de cirage (échantillon *Gu*) que chez les petits.

b. Le test de comparaison appliqué suppose que les variances des deux échantillons sont égales et pour les comparer validement, il faut que la distribution de la variable étudiée (l'appréciation ici) soit normale. Or, la valeur standardisée des statistiques de forme montrent une importante distorsion (symétrie ou aplatissement selon le groupe. Théoriquement, le test paramétrique devrait être remplacé par un test non paramétrique. Dans les faits, comme l'intervalle de confiance d'une moyenne est robuste et n'est pas sensible au non-respect de l'hypothèse de normalité, de manière pragmatique, c'est cette méthode qui utilisée. Le paragraphe suivant développe néanmoins la résolution de cette situation 4 avec les deux tests non-paramétriques les plus usuels.

4.4 Formulation des hypothèses et des conclusions avec un test non paramétrique

Deux tests sont applicables : celui de Mann-Whitney compare les médianes et intègre la dimension ordinale des valeurs ; et celui de Kolmogorov-Smirnov compare la distribution des valeurs contenues dans les deux échantillons étudiés.

Le Test de Mann-Whitney

Ce test est calculé en combinant les deux échantillons étudiés, où les données sont triées de la plus petite à la plus grande valeur. Il repose sur la comparaison des rangs moyens des deux échantillons dans les données combinées.

- Formulation des hypothèses :
  - hypothèse nulle H_0 : médiane de *Pu* – médiane de *Gu* = 0
  - hypothèse alternative H_1 : médiane de *Pu* – médiane de *Gu* ≠ 0
  - avec un seuil de risque α de 5 % (soit un seuil de confiance de 95 %)

© Dunod – Toute reproduction non autorisée est un délit.

- Résultats du test :
  - rang moyen de l'échantillon *Pu* : 175,45
  - rang moyen de l'échantillon *Gu* : 232,07
  - valeur critique de la statistique W = 26408,0
  - probabilité critique α associée à W = 0,00000142164 (très inférieure à 0,05)

Conclusion *sur la base du test de Mann-Whitney* : la valeur de la probabilité critique étant très inférieure à 0,05, il y a une différence statistiquement significative entre les médianes (au niveau de confiance de 95,0 %) : comme le rang moyen des *Gu* est supérieur à celui des *Pu*, on peut conclure que les gros utilisateurs ont une meilleure appréciation de l'applicateur testé.

Le Test de Kolmogorov-Smirnov

Ce test de comparaison des distributions est fondé sur le calcul de la distance maximale qui existe entre les distributions cumulées des deux échantillons testés.

- Formulation des hypothèses :
  - hypothèse nulle H_0 : Distribution *Pu* = Distribution *Gu*
  - hypothèse alternative H_1 : Distribution *Pu* ≠ Distribution *Gu*
  - avec un seuil de risque α de 5 % (soit un seuil de confiance de 95 %)
- Résultats du test :
  - valeur critique de la statistique DN (distance maximale) = 0,492
  - valeur de la statistique K-S bilatérale = 4,899
  - probabilité critique α approximative associée = 0,00 (inférieure à 0,05)

Conclusion *sur la base du test de Kolmogorov-Smirnov* : puisque la probabilité est inférieure à 0,05 (5 %), il y a une différence statistiquement significative entre les deux distributions observées : elles n'appartiennent donc pas à la même population de référence. La moyenne des gros utilisateurs étant la plus élevée, on peut dire qu'ils apprécient mieux que les autres les performances de l'applicateur testé.

5 La comparaison entre deux échantillons appariés – le cas de la moyenne (décision 5)

5.1 Problématique décisionnelle et type de test statistique

Le chef de produit veut vérifier que sa nouvelle offre possède bien un avantage concurrentiel. Si c'est le cas, son évaluation doit être *au moins* équivalente, sinon supérieure, à celle du produit concurrent usuel. Or, ce dernier a été évalué lors de la première visite et le tube en test lors de la seconde visite *par les mêmes personnes*.

De ce fait, son évaluation n'est pas indépendante de celle du produit usuel qui prend alors valeur de référence : le tube s'y trouve implicitement comparé.

Cette situation relève de la *comparaison d'une moyenne entre deux échantillons de mesures appariés*. Techniquement, pour confirmer l'avantage concurrentiel espéré, il suffit de vérifier que la moyenne de la variable construite – résultant de la différence calculée entre les deux mesures effectuées sur chaque *même* personne de l'échantillon – est significativement supérieure à 0 (et non pas comparer la moyenne de chacune des deux séries de mesures comme cela se ferait dans le cas de deux échantillons indépendants).

5.2 Formulation des hypothèses et résultats des tests

Données	
Effectif des paires comparées : $n = 120$; Moyenne $\bar{D}$ des différences des 120 observations : $\bar{D} = -0,267$; écart-type : $s_D = 1,128$; aplatissement standardisé : - 0,895 ; asymétrie standardisée : - 0,600	
Hypothèses nulle et alternative *(test unilatéral gauche)*	
On note D la variable aléatoire obtenue comme étant la différence observée entre les deux mesures effectuées (produit usuel [*U*] et en test [*T*]) : $D = X_U - X_T$ $H_0 : \mu_U = \mu_T$ donc $\mu_D = \mu_U - \mu_T = 0$ (comprendre $\mu_D \geq 0$) $H_1 : \mu_U < \mu_T$ donc $\mu_D = \mu_U - \mu_T < 0$	
Intervalle de confiance de la différence au seuil α : 0,05	
$\bar{D} \pm z_{\alpha/2} \frac{s_D}{\sqrt{n}}$	$-0,267 \pm 1,96 \frac{1,128}{\sqrt{120}}$ $-0,267 \pm 1,96 \times 0,103 = [-0,469 ; -0,065]$
Test de comparaison	
$z = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}}$	$z = \frac{-0,267}{\frac{1,128}{\sqrt{120}}} = -2,593$ probabilité critique associée : 0,00476 < 0,05
Test non paramétrique *fondé sur la médiane :* test du signe	
Ho : $\text{médiane}_{différences} = 0$ H_1 : $\text{médiane}_{différences} \neq 0$	Nombre de valeurs < médiane hypothétique : 52 Nombre de valeurs > médiane hypothétique : 33 Valeur du test : 1,9524 ; Probabilité critique : 0,0509

Pour savoir si l'offre testée est perçue au moins aussi bonne que le produit usuel, il faut vérifier si la moyenne des écarts entre les paires d'observations est différente de 0 significativement.

Par hypothèse, il n'y a pas d'écart de mesures au sein de chaque paire : la somme des écarts étant nulle, la moyenne des écarts vaut donc zéro [0]. Or, le résultat des tests montre que :

1. la différence théorique [0], qui correspond à l'hypothèse nulle, n'est pas incluse dans l'intervalle de confiance de la moyenne des différences entre les paires

© Dunod – Toute reproduction non autorisée est un délit.

d'observations dont les limites sont comprises entre -0,469 et -0,065 (au seuil de confiance de 95 %) ;

2. la valeur absolue du test t (-| 2,593 |) est supérieure à sa valeur théorique (1,96) ;
3. la probabilité critique associée à cette valeur du test (0,0095 = 0,95 %) est très inférieure au seuil de risque fixé de 5 %.

Ces résultats conduisent à rejeter H_0 pour admettre que H_1 est plausible : le produit en test est perçu significativement supérieur au produit habituellement utilisé : le signe « moins » associé à la valeur du t-test indique que, en moyenne, les valeurs de la première mesure (produit usuel) sont inférieures à celle de la seconde mesure (produit testé).

5.3 Formulation des conclusions et recommandations

Puisque l'hypothèse alternative H1 est plausible, on peut admettre que l'offre testée avec son tube applicateur a une réelle valeur concurrentielle : en moyenne, il est considéré comme supérieur au produit habituellement utilisé. Rien ne s'oppose donc à un lancement centré sur sa qualité reconnue.

5.4 Formulation des hypothèses et des conclusions avec un test non paramétrique : le test du signe

L'examen des caractéristiques de la variable mesurant la différence entre les paires d'observations montre qu'elle suit une loi normale (indices de forme dans la norme : ± 2), ce qui justifie l'usage d'un test paramétrique. À défaut, *le test du signe* permettrait de résoudre le même problème. Il a l'avantage d'être moins sensible aux éventuelles valeurs extrêmes que la variable mesurant la différence.

Le test du signe repose sur la médiane de la variable mesurant la différence entre les paires d'observations. Il est fondé sur le *comptage du nombre de valeurs* situées au-dessus et au-dessous de la médiane hypothétique. On note que son résultat est à la limite de la significativité : .0509 > .05 (*cf.* tableau ci-dessus). En toute rigueur, l'utilisation de ce test, moins puissant que le test t fondé sur la loi normale, impliquerait d'admettre l'hypothèse nulle et que le tube applicateur testé n'est pas perçu significativement plus performant que le produit habituellement utilisé.

Comme c'est le nombre de valeurs inférieures à la médiane hypothétique ($médiane_U$ – $médiane_T$) qui est le plus élevé (52), l'écart est donc en faveur du produit testé : le nouveau tube n'est donc pas de moins performant ; mais cette affirmation n'est pas soutenue par le résultat du test statistique.

6 La comparaison de rangs entre deux échantillons indépendants (décision 6)

6.1 Problématique décisionnelle et type de test

Le décideur se demande si l'*importance* accordée au soin des cuirs diffère selon les hommes et les femmes, car cet élément pourrait être intégré dans la conception des stratégies publicitaires. Techniquement, il s'agit de comparer la distribution des réponses obtenues sur deux échantillons indépendants sur une *variable ordinale*.

En fait, cette procédure met en relation deux variables non métriques : l'importance accordée à l'entretien des cuirs (variable ordinale) et l'appartenance au genre (variable nominale). Si la relation est avérée, l'importance déclarée pour le soin du cuir diffère donc selon le sexe du répondant.

Le traitement d'une telle relation bivariée peut être envisagé de deux manières :
- Soit l'analyse de *la hiérarchie ne s'impose pas* (la variable « importance » est considérée comme nominale et non ordinale). À la place d'un test de comparaison entre la distribution des réponses de deux échantillons, il convient plutôt de tester le lien entre les deux variables avec un *test d'indépendance* fondé sur la distribution du Chi-deux.
- Soit l'analyse de *la hiérarchie s'impose* : il faut utiliser un test qui permet de tenir compte du rang en testant la différence entre la médiane de chaque échantillon pour déterminer lequel a le meilleur rang. Trois tests sont disponibles : le test de Kolmogorov-Smirnov, le test de la médiane et le test U de Mann-Whitney.

6.2 Mise en œuvre des différentes procédures statistiques

Le test d'indépendance par le Chi-deux

Ce test postule l'indépendance de la distribution des observations entre les modalités de la variable en lignes et celles de la variable en colonnes. Cette indépendance signifie qu'il n'y a pas de lien entre les deux variables (équivalent de l'hypothèse nulle) ; en cas de rejet de l'hypothèse d'indépendance (équivalent de l'hypothèse alternative), la distribution des valeurs d'une variable dépend de la distribution des valeurs de l'autre variable[1].

Formulation du test d'hypothèse

Il s'agit de tester l'indépendance entre la distribution des valeurs de la variable *Importance du soin* en fonction des modalités de la variable *Sexe*. L'hypothèse nulle

1. Des indicateurs statistiques spécifiques, comme le coefficient Lambda de Guttman permettent de déterminer l'utilité d'un facteur (ligne ou colonne) pour prévoir la distribution entre les modalités de l'autre facteur *(cf.* chapitre 3, section 1, § 3.2).

© Dunod – Toute reproduction non autorisée est un délit.

postule que la distribution du tableau de contingence entre les deux variables est identique à celui construit sous l'hypothèse d'indépendance. Le test est effectué avec un seuil de risque α de 5 % (soit un seuil de confiance de 95 %).

Résultats du test

Le Chi-deux (χ^2) calculé vaut 18,66 alors que sa valeur critique pour 2 degrés de liberté vaut 5,991 et la probabilité critique associée : 0,0001 < 0,05 (seuil usuel). Il faut donc rejeter l'hypothèse de l'indépendance entre les deux variables.

Interprétation du résultat statistique.

Le rejet de l'hypothèse d'indépendance conduit à admettre que les deux variables *Sexe* et *Importance de l'entretien* sont liées. L'analyse détaillée de la décomposition des valeurs du Chi-deux entre les différentes cellules issues du croisement des deux variables (*cf.* section 1 du ch. 3) montre que les femmes attachent plus d'importance au soin des articles en cuirs que les hommes.

Le test U de Mann-Whitney

Ce test compare les médianes des deux échantillons. Après avoir combiné les deux échantillons, il consiste à réunir et à trier les observations des deux échantillons de la plus petite à la plus grande, puis à comparer leurs rangs moyens.

Formulation du test d'hypothèse

L'hypothèse nulle postule que les deux échantillons sont positionnés de manière identique (les médianes sont identiques).

$H_o : Mé_H = Mé_F$

$H_1 : Mé_H \neq Mé_F$

Résultats du test

Le rang moyen pour l'échantillon d'Hommes est 104,5 alors que celui des Femmes est 96,5 ; en moyenne, les Femmes sont « mieux classées » que les Hommes (96,5 < 104,5) – rappelons que si la variable est ordinale, le classement le meilleur est celui le plus proche de 1 (*cf.* chapitre 1). La valeur critique du test (U) calculée est 4 600 ; la probabilité critique α associée : 0,294 > 0,05, seuil de risque α fixé. On accepte alors l'hypothèse d'égalité entre les médianes des deux échantillons.

Interprétation du résultat statistique

L'acceptation de l'hypothèse nulle avec ce test conduit à admettre que les hommes et les femmes de l'échantillon n'ont pas un niveau d'exigence différent à l'égard de l'entretien des articles en cuirs : l'écart de rang observé n'est pas significatif.

Le test de Kolmogorov-Smirnov

Ce test de comparaison des distributions est fondé sur le calcul de la distance maximale qui existe entre les distributions cumulées des deux échantillons testés.

Formulation des hypothèses

L'hypothèse nulle postule que les distributions sont identiques (H_0 : distribution des Hommes = distribution des Femmes) et l'hypothèse alternative, qu'elles sont différentes (H_1 : distribution des Hommes ≠ distribution des Femmes). Le test s'effectue avec un seuil de risque α de 5 % (soit un seuil de confiance de 95 %).

Résultats du test

La valeur critique de la statistique DN (distance maximale) est 0,50 alors que la valeur de la statistique K-S bilatérale est 3,536. La probabilité critique α associée vaut approximativement = 0,00 (inférieure à 0,05).

Interprétation du résultat statistique

La valeur de la probabilité étant inférieure à 0,05 (5 %), il y a donc une différence statistiquement significative dans la distribution de l'importance de l'entretien entre les deux échantillons. La comparaison des distributions des réponses montre que ce sont les Femmes les plus nombreuses à se montrer le plus sensible à l'entretien des cuirs (modalité 1 « très important » : 40 % contre 20 %).

Cet exemple montre que tous les tests ne conduisent pas aux mêmes résultats, car ils ne reposent pas sur les mêmes paramètres d'analyse. La réflexion préalable au choix du test le mieux adapté à la situation décisionnelle est donc déterminante.

7 La comparaison simultanée de plusieurs échantillons (décision 7)

Il y a deux procédures pour comparer plusieurs échantillons simultanément.

La première, très analytique, consiste à comparer chaque paire d'échantillons de manière systématique avec le test adapté choisi parmi ceux qui viennent d'être présentés. Il suffit ensuite d'ordonner les échantillons en fonction de la significativité des résultats pour tirer les conclusions opérationnelles qui s'imposent.

La seconde méthode, plus globale, repose sur le principe de *l'analyse de la variance*. Elle consiste à vérifier si au moins un échantillon diffère significativement des autres et, dans l'affirmative, à n'opérer qu'ensuite les comparaisons entre les paires d'échantillons. Selon la nature des variables étudiées, on choisit entre l'analyse de variance métrique (test de Fisher-Snedecor) et l'analyse de variance non métrique (test de Kruskal-Wallis).

© Dunod – Toute reproduction non autorisée est un délit.

7.1 Le principe d'une analyse de variance : des données au résultat

La comparaison de plusieurs échantillons correspond à une analyse qui croise deux variables ayant chacune un statut propre (analyse bivariée). Une variable définit les échantillons : c'est le plus souvent une variable non métrique appelé *facteur* utilisée pour caractériser les observations. Dans la décision 7, les trois argumentaires correspondent aux trois modalités d'une variable nominale. Une autre variable définit la mesure effectuée qui est nécessairement métrique pour le test de Fisher-Snedecor ou ordinale pour le test de Kruskal-Wallis. Dans l'exemple, la variable composite résultant de la combinaison des deux critères (crédibilité et incitation) est métrique : sa moyenne est calculée par type d'argument.

La comparaison des bornes de l'intervalle de confiance de chaque moyenne (au seuil de risque usuel de 5 %) montre qu'elles se chevauchent pour les arguments 2 et 3 : ils ne sont donc pas significativement différents et opèrent avec la même efficacité. La moyenne du premier argumentaire a des bornes extérieures à celles des deux autres : il s'en distingue donc. On peut ainsi rapidement conclure que les arguments 2 et 3 sont d'efficacité équivalente et significativement supérieure au premier argument ; on peut décider d'abandonner l'argument 1, mais lequel choisir entre 2 et 3 ? L'analyse de variance permet d'avoir une approche plus précise.

Logique de l'analyse de variance métrique : test de Fisher-Snedecor

Le modèle d'analyse de variance métrique consiste à tester simultanément l'homogénéité des moyennes de plusieurs échantillons. Soit échantillons indépendants de tailles (identiques ou non) n_k issus de manière aléatoire d'une même population et suivant une loi normale de variance identique : le test de Fisher permet de comparer les k moyennes de ces populations μ_1, μ_2,…, μ_k pour un seuil d'erreur donné :

– l'*hypothèse nulle* du test postule que toutes les moyennes sont identiques :

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

– l'*hypothèse alternative* postule qu'au moins deux moyennes sont significativement différentes l'une de l'autre :

$$H_1 : \mu_i \neq \mu_j \ (i \neq j)$$

Ce modèle d'analyse décompose la variance totale en deux sources :

– celle due aux aléas d'échantillonnage, appelée *variance intra-groupe* ou *variance résiduelle* (V_R) ;

– celle due à une véritable différence de structure entre les échantillons, appelée *variance inter-groupe* ou *variance factorielle* (V_F) parce que ces échantillons sont la source, le facteur déclenchant des différences.

La statistique du test n'est autre que le ratio $F_{obs} = V_F/V_R$ qui mesure l'importance de la variance inter-groupe V_F par rapport à la variance intra-groupe V_R. C'est cette statistique qui est utilisée pour tester l'homogénéité des moyennes. La variable

aléatoire F est un rapport de variances qui suit la Loi de Fisher-Snedecor à $v_1 = k - 1$, $v_2 = n - k$ degrés de liberté. Selon la valeur de ce ratio :

- Si F<1 ($V_F < V_R$), la variance résiduelle (V_R) est plus grande que la variance factorielle (V_F), alors on est assuré que l'hypothèse nulle est vraie (pas de différence entre les moyennes) ;
- Si F=1 ($V_F = V_R$), les moyennes sont identiques (décision en faveur de H_0) ;
- Si F>1 ($V_F > V_R$), au moins la moyenne d'un des échantillons se différencie significativement des autres, mais à la condition que la valeur calculée F_{obs} soit supérieure à la valeur critique théorique $F_{\alpha|v1|v2}$ pour un seuil de risque α fixé : dans ce cas, l'hypothèse nulle d'égalité des moyennes est rejetée.

Le test de Fisher ne précise pas quelle moyenne est statistiquement différente de n'importe laquelle des autres. Pour cela, il faut recourir à l'un des tests *post hoc* qui existent : tests *de Tukey*, *de Newman-Keuls*, *de Duncan*, *de Scheffé, de Dunnett*, *de Bonferroni*, etc. dont la présentation théorique dépasse le cadre de cet ouvrage.

Remarque

Pour un seuil de risque fixé, les valeurs de F se distribuent selon une loi de distribution tabulée par Fisher-Snedecor à k-1 degrés de liberté pour le numérateur (V_F) – où k est le nombre d'échantillons comparés – et à n-k degrés de liberté pour le dénominateur (V_R) où *n* est l'effectif total d'échantillons comparés. Dans la décision 7 : n = 331 [98+124+109] et k = 3, soit les deux degrés de liberté : k-1 = 3-1 = 2 et n-k = 331-3 = 328.

L'analyse de variance métrique (modèle de Fisher-Snedecor) implique une distribution normale de la variable métrique et une homogénéité des variances des échantillons (différents tests permettent de le vérifier). À défaut, *il faut utiliser* le modèle non métrique de Kruskal-Wallis qui repose sur l'analyse des médianes.

Focus

Tester l'égalité des variances des échantillons : homoscédasticité

Si tous les échantillons ont la même taille, deux tests simples permettent de vérifier l'homogénéité de la variance des échantillons : le test de Hartley ou le test de Cochran. Dans le cas contraire, il est préférable d'utiliser le test de Bartlett ou le test de Levene.

Test de Hartley

Le test de Hartley consiste à calculer le rapport entre la plus grande variance corrigée et la plus petite : $F_{max} = s^2_{max}/s^2_{min}$. On rejette l'hypothèse de l'égalité des variances si la valeur calculée est supérieure au seuil critique donné par la table de Hartley.

Test de Cochran

Ce test est une variante du précédent. La statistique C du test est le rapport entre la plus grande variance corrigée et la somme de la variance de tous les échantillons :

☞

© Dunod – Toute reproduction non autorisée est un délit.

☞

$F_{max} = s^2_{max} / \sum_{i=1}^{k} s_i^2$. Comme pour le test de Hartley, la statistique C est comparée avec le seuil critique C_α donné par la table de Hartley.

Test de Bartlett

Il consiste à calculer une statistique qui dépend du nombre d'échantillons k, de la taille de chacun et de l'estimation globale de la variance. La statistique suit une loi du χ^2 à $k - 1$ degrés de liberté. L'hypothèse nulle est alors rejetée si la valeur observée du χ^2 est supérieure à la valeur critique limite supérieure de la distribution $\chi^2_{k-1|\alpha}$

Test de Levene

Le test de Levene consiste à analyser les écarts absolus des valeurs aux moyennes des échantillons. Il est basé sur l'hypothèse que si les variances des échantillons sont proches, les écarts absolus moyens des échantillons ne sont pas significativement différents. Il utilise la procédure d'analyse de la variance dont la complexité ne sera pas développée ici. Il existe une variante de ce test - le test de Brown & Forsythe - basé sur l'analyse des écarts entre médianes au lieu des moyennes. Les tests de Levene et de Brown & Forsythe sont moins sensibles à la contrainte de normalité des échantillons.

Logique de l'analyse de variance non-métrique : test de Kruskal-Wallis

Extension du test de Mann-Whitney appliqué à plusieurs échantillons comparés simultanément, le test de Kruskal-Wallis formule l'hypothèse nulle que les médianes des échantillons étudiés sont les mêmes. Les données des différents échantillons sont triées de la plus petite à la plus grande en déterminant leur rang dont la somme est calculée pour chaque échantillon.

La statistique H de Kruskal-Wallis est calculée en fonction de la somme des rangs des échantillons. Si la taille de chaque échantillon est supérieure à 5 observations, H se distribue selon un Chi-deux à k-1 degrés de liberté (où k est le nombre d'échantillons). Si la statistique H est supérieure à la valeur du Chi-deux, l'hypothèse nulle de l'égalité des échantillons est rejetée au bénéfice de l'hypothèse alternative : au moins un échantillon est différent des autres. Quand la taille de l'échantillon est inférieure à 5 observations (ou l'ensemble inférieur à 15 observations), H a sa propre loi de distribution qui est tabulée.

Remarque

La variable non métrique constituant les échantillons comparés identifie la source de variation des moyennes. Quand cette source est volontairement contrôlée par l'analyste, ce modèle peut être utilisé comme outil d'analyse de la causalité pour mesurer l'effet ou l'influence d'une variable sur une autre variable. Les plans d'expérience, qui sont des plans d'échantillonnage spécifiques, sont utilisés pour contrôler les sources de variation, les « effets » de la variable constituant les échantillons comparés.

7.2 Problématique et résolution avec une ANOVA

L'objectif est d'identifier, parmi les trois argumentaires conçus, celui susceptible d'être le plus efficace face aux offres concurrentes. Il faut donc savoir si l'un d'entre eux se montre particulièrement convaincant sur la base du critère d'efficacité utilisé. L'analyse de variance est le modèle adapté à ce problème puisqu'il permet de comparer simultanément les moyennes des trois argumentaires.

Quand la variable mesurée dans chaque échantillon est métrique, le modèle d'analyse de variance s'impose. Il est souvent dénommé « ANOVA » qui est l'abréviation de l'expression anglaise *An*alysis *of Va*riance, aussi appelé « analyse de variance univariée », car ne traitant qu'une seule variable métrique à la fois.

Tableau 2.3 – Application à l'exemple du choix d'un argumentaire publicitaire

Paramètres	Statistiques
Argumentaire 1	n = 98 ; moyenne = 3,65 ; écart-type = 0,943 ; asymétrie = 3,037 ; aplatissement = - 2,962
Argumentaire 2	n = 124 ; moyenne = 6,50 ; écart-type = 1,108 ; asymétrie = 0 ; aplatissement = - 3,026
Argumentaire 3	n = 109 ; moyenne = 5,91 ; écart-type = 1,512 ; asymétrie : 0,605 ; aplatissement : - 2,738
Homogénéité des variances :	Test C de Cochran : 0,519 Proba. = 0,0000 < 0,05 Test de Bartlett : 1,078 Proba. = 0,0000 < 0,05 La contrainte d'égalité des variances n'est pas satisfaite
Test F de Fisher-Snedecor :	F : 161,19 ; probabilité critique : 0,0000 ; rejet de H_0
Test H de Kruskal-Wallis :	H : 165,26 ; probabilité critique : 0,0 ; rejet de H_0

7.3 Formulation des hypothèses, résultat des tests et conclusions

L'hypothèse nulle pose que les écarts observés entre les moyennes des trois argumentaires ne sont dus qu'aux aléas d'échantillonnage et ne dénotent pas une réelle différence d'efficacité. L'hypothèse alternative pose qu'au moins une de ces moyennes est différente d'une autre.

Les résultats d'une analyse de variance se présentent sous la forme d'un tableau comprenant les résultats de l'analyse statistique (le test F) et la comparaison systématique des moyennes de chaque groupe étudié.

La variance inter-groupe (V_F) est nettement supérieure à la variance résiduelle (V_R) et le ratio $F = V_F/V_R$ est hautement significatif (probabilité d'observer la valeur 169,19 inférieure à 1 chance sur 10 000). Il faut rejeter l'hypothèse H_0 d'égalité des moyennes : au moins une moyenne se différencie significativement des autres. Pour savoir quelle est la moyenne significativement différente, une information complémentaire est fournie – appelée « contraste » – qui compare systématiquement les moyennes par paires en fournissant la significativité des écarts constatés.

© Dunod – Toute reproduction non autorisée est un délit.

Tableau 2.4 – Exemple des résultats d'une analyse de variance paramétrique

Analyse de variance*					
Source de Variation	**Somme des carrés**	**Ddl**	**Carré moyen**	**F**	**Probabilité α**
Inter-groupes (V_F)	475.991	2	237.996	161.19	0.0000
Intra-groupes (V_r)	484.287	328	1.47648		
Total (Corr.)	960.278	330			

Tests des étendues multiples de la variable Note selon chaque Argumentaire					
Méthode : pour un seuil de confiance de 95.0 %, avec la méthode de Scheffé :					
n° argument	**Effectif**	**Moyenne**	**Groupes homogènes**		
1	98	3.653	X		
3	109	5.908		X	
2	124	6.500			X
Contraste			**Différence**		**+/- limites**
1 - 2			*-2.84694		0.40385
1 - 3			*-2.2552		0.415936
2 - 3			* 0.59174		0.392302
*indique une différence statistiquement significative.					

* tableau de résultats obtenu sous Statgraphics

Remarque

Différents tests (Tuckey, Bonferri, Duncan…) permettent d'effectuer ces comparaisons dont les logiciels précisent les conditions d'usage. La méthode de Scheffé est adaptée quand les échantillons ne résultent pas d'un plan d'expérience ; la méthode de Bonferri est utilisée si la statistique F n'est pas significative, mais que l'analyste veut connaître la hiérarchie des écarts dans le cadre d'échantillons issus d'un plan d'expérience.

Pour la décision 7, le test montre que les trois moyennes sont significativement différentes : les trois argumentaires peuvent donc être hiérarchisés en fonction de leur capacité à convaincre, respectivement le second (6,50), puis le troisième (5,91) et, peu performant, le premier testé (3,65). En conséquence, le responsable devra éviter le premier argumentaire mais pourra choisir l'un des deux plus performants entre lesquels l'écart est faible bien que significatif.

En toute rigueur, l'analyse de variance métrique ne devrait pas être utilisée sur ces données (mais plutôt le modèle de Kruskal-Wallis), car ni la variable ne se distribue normalement (indices de forme déviants), ni l'égalité des variances n'est respectée. Cependant ce test est robuste, même quand les spécifications théoriques ne sont pas toutes respectées.

7.4 Résolution avec un modèle non-métrique : le test de Kruskal-Wallis

La problématique reste la même. Mais ce test compare les rangs moyens calculés dans chaque échantillon.

Tableau 2.5 – Exemple des résultats d'une analyse de variance non paramétrique

Test de Kruskal-Wallis pour la variable Note selon chaque argumentaire		
n° d'argumentaire	**Taille d'échantillon**	**Rang moyen**
1	98	65,827
2	124	223,573
3	109	190,569
Test statistique : H = 165,263 Probabilité critique = 0,0		

Calcul de la statistique H de Kruskal-Wallis

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{K} \frac{R_j^2}{n_j} - 3(N+1)$$ Où :

- N : nombre total d'observations (tous échantillons confondus)
- n_j : nombre d'observations dans un échantillon
- K : nombre d'échantillons
- R_j : somme des rangs dans un échantillon

Puisque la valeur de la statistique H calculée est très significative, l'hypothèse d'égalité des rangs dans les échantillons ne peut être retenue. Les échantillons se distinguent donc bien et l'analyse des rangs moyens permet de hiérarchiser les trois argumentaires. Les conclusions sont identiques à celles de l'analyse de variance : le second argumentaire s'impose et le premier est à éviter, car très mal classé.

L'ESSENTIEL

■ *Une méthodologie pour construire la comparaison*

Envisager de prendre une décision managériale sur la base de la comparaison de variables susceptibles de décrire des échantillons implique de respecter une démarche méthodologique liée à l'analyse du contexte de la décision et au respect d'un certain nombre de contraintes dépendantes des outils statistiques :

- La nature métrique ou non de la variable sur laquelle repose la comparaison oriente le choix vers une méthode statistique de comparaison.
- Le nombre et la nature – indépendante ou appariée – des échantillons de mesure comparés définissent le type d'outil statistique adapté : comparaison directe des moyennes ou estimation de l'écart moyen comme valeur nulle.
- La problématique décisionnelle détermine quelle est l'hypothèse statistique à vérifier et précise comment l'hypothèse alternative doit être formulée.

☞

© Dunod – Toute reproduction non autorisée est un délit.

☞

– La maîtrise, indispensable, des conditions d'interprétation des résultats à l'aide de la valeur critique ou de la probabilité critique est garante d'une bonne interprétation du résultat statistique, nécessaire pour une prise de décision managériale réfléchie.
– L'essentiel n'est pas dans la maîtrise des calculs – les algorithmes les font très bien – mais dans la maîtrise des conditions d'utilisation des outils disponibles et de la capacité de passer du résultat statistique à la conclusion managériale et opérationnelle.

Chapitre 3

Comment exploiter un tableau de contingence

OBJECTIFS

- Comprendre la logique d'un test d'indépendance et savoir interpréter le résultat statistique pour l'utiliser de manière utile et pertinente dans la prise de décision.
- Exploiter un tableau de contingence avec l'analyse des correspondances simples et en maîtriser l'usage : de la construction de tableau de données à l'exploitation des résultats statistiques en évitant toute interprétation abusive.

SOMMAIRE

SECTION 1 Exploiter un tri croisé

SECTION 2 L'analyse des correspondances pour exploiter un tableau de contingence

Une décision implique souvent de considérer plusieurs variables simultanément. Ainsi, nombre de données dont dispose le gestionnaire se présente sous la forme d'un tableau à double entrée croisant deux variables et dans lequel chaque cellule contient soit l'effectif, soit la fréquence des modalités associées par paires : ce sont les tableaux croisés. Leur analyse permet d'évaluer dans quelle mesure deux variables non-métriques (ou métriques discrétisées) sont liées ou, au contraire, sont indépendantes.

> En marketing : vérifier si les intentions d'usage et d'achat d'une lingette pour dépoussiérer varient en fonction de la sensibilité à la propreté, de la pratique du ménage, du niveau de scolarisation, de l'activité professionnelle, etc. ; contrôler, suite à diverses actions de communication (publicitaire et promotionnelle) qu'une marque de produits cosmétiques est reconnue pour posséder mieux que ses concurrentes certains attributs, particulièrement différenciant et commercialement motivants. En ressources humaines : analyser s'il existe une différence significative entre le comportement des hommes et celui des femmes ; étudier les liens qui existent entre le type de stratégie, les types de structure, les styles de management, les savoir-faire, les valeurs, etc.

L'étude de l'association entre deux variables – objet de la première section de ce chapitre – peut être abordée dans une triple perspective :

– Analyser la *simple association* entre deux variables qualitatives. Par exemple : la lecture de la presse quotidienne et l'audience des chaînes d'information TV sont-elles liées ? Ces deux variables sont symétriques : l'audience de l'une ne dépend pas *a priori* de l'audience de l'autre. Un test d'indépendance, tel que le test du Chi-deux (χ^2), permet de vérifier s'il y a ou non un lien. Deux autres indices statistiques (les coefficients C de contingence et V de Cramer) permettent de mesurer l'intensité d'une éventuelle association, ce dont le Chi-deux ne peut rendre compte.
– Analyser une *relation hiérarchique* supposée qui peut être interprétée comme une relation causale. Par exemple, la lecture de la presse quotidienne est-elle liée à l'activité professionnelle ? Plus ou moins explicitement, on fait l'hypothèse que la lecture de la presse quotidienne peut varier selon la profession exercée ; en d'autres termes : il y a une segmentation des titres lus en fonction de la profession. Dans cette logique, les deux variables sont hiérarchisées et ne partagent pas le même statut : comme la profession exercée détermine la lecture des titres de presse, alors cette variable est considérée comme causale (bien qu'elle ne le soit pas sur le plan méthodologique). Le test repose bien sur le Chi-deux mais le tableau croisé sera construit de telle manière que la variable dépendante (la lecture de la presse ici) apparaisse bien comme étant sous l'influence de la variable indépendante (la profession). Une analyse des Chi-deux partiels (part du Chi-deux dans chaque cellule) est alors nécessaire pour aider à comprendre le sens de l'effet supposé.
– Donner à cette perspective hiérarchique une *dimension prédictive*. Dès lors qu'une relation de dépendance est supposée, il est logique de chercher à prévoir les valeurs de la variable dépendante à partir des variations de la variable indépen-

dante. Le coefficient lambda (Λ) de Guttman contribue à cette évaluation. Ce coefficient est aussi un moyen pour vérifier *a posteriori* la pertinence de l'hypothèse d'une relation hiérarchique entre les deux variables étudiées.

L'étude des liens entre deux variables non-métriques est assez facile quand le nombre de leurs modalités reste limité, mais dès qu'il augmente, une analyse directe devient plus difficile. Or, dans nombre de cas, les deux variables croisées comportent chacune un nombre conséquent de modalités.

Pour analyser ce type de tableau croisé – appelé aussi *tableau de contingence* – il existe une procédure statistique, l'*analyse des correspondances*, dérivée du Chi-deux. Cette méthode est d'un usage très courant en marketing, par exemple pour étudier le positionnement de marques sur la base d'une série d'items. Les informations de base – marques et items – sont alors transformées pour constituer chacune une variable spécifique : les différentes marques étudiées constituent les modalités de la variable « marque » et les différents items utilisés pour la comparaison : les modalités « image ».

L'analyse des correspondances permet de vérifier si les deux variables étudiées sont ou non liées et, si elles le sont significativement, d'identifier quelles sont les modalités associées, non seulement entre les deux variables, mais aussi *au sein de chaque variable*. La seconde section de ce chapitre est entièrement consacrée à la technique d'analyse des correspondances simples.

Section 1 EXPLOITER UN TRI CROISÉ

Cas d'illustration

Choix d'un type de train par les hommes et les femmes

Dans l'établissement de ses statistiques de fréquentation sur une relation ferroviaire, la SNCF constate des différences de population selon le type de train : trains du matin, de mi-journée, de début de soirée et de nuit. Au vu de ses statistiques, elle se demande s'il n'existerait pas une relation significative entre le type de train fréquenté et le sexe du voyageur.

Cette question sous-entend en fait une hypothèse : selon le sexe, il se pourrait qu'on n'ait pas l'habitude de fréquenter le même type de train, et donc, si l'hypothèse est avérée, que le sexe détermine le choix d'un type de train plutôt qu'un autre. Cela pourrait alors avoir une incidence sur l'offre de services que la SNCF pourrait envisager, principalement sur les trains de nuit.

☞

© Dunod – Toute reproduction non autorisée est un délit.

☞ À cette fin, une enquête a été conduite auprès d'un échantillon de passagers sélectionnés de manière aléatoire sur la relation étudiée afin de recueillir, entre autres choses, leurs habitudes déclarées concernant le choix d'un horaire de train quand ils doivent faire ce trajet. L'étude porte sur un échantillon de 125 personnes qui effectuent ce trajet au minimum une fois par mois, dont le tableau ci-dessous donne la répartition des effectifs (les chiffres entre parenthèses sont obtenus en divisant les effectifs de toutes les cellules du tableau de contingence par les effectifs globaux qui expriment les effectifs en pourcentage).

	« Habitude de choix d'un type de train »				
« Genre »	1- matin	2- journée	3- soirée	4- nuit	***Total***
1 - Féminin	10 (8 %)	9 (7,2 %)	19 (15,2 %)	7 (5,6 %)	***45*** **(36 %)**
2 - Masculin	20 (16 %)	12 (9,6 %)	16 (12,8 %)	32 (25,6 %)	***80*** **(64 %)**
Total	***30*** **(24 %)**	***21*** **(16,8 %)**	***35*** **(28 %)**	***39*** **(31,2 %)**	***125*** **(100 %)**

1 Construire le tableau croisé en fonction de la problématique

Dans ce cas d'illustration, en passant de la préoccupation de l'analyste à celle du statisticien, la question revient à se demander si le choix d'un train dépend du sexe du voyageur. Cette question s'inscrit dans une perspective explicative : en toute logique, c'est bien le sexe qui détermine le choix d'un train et non l'inverse.

Sur le plan technique, il s'agit de vérifier l'existence d'un lien entre la variable « sexe » (deux modalités : 1- féminin et 2- masculin) et la variable « type de train » (quatre modalités, pour trains : 1- du matin, 2- de milieu de journée, 3- de fin de journée et 4- de nuit). Mais la problématique, telle que formulée dans cette perspective causale, organise la disposition du tableau croisé : situation des variables et calcul des pourcentages.

1.1 Lecture et interprétation des différents pourcentages calculables

Dans un tableau croisé, le calcul des pourcentages relatif au nombre d'observations peut être effectué sur trois bases différentes : le total, les lignes ou les colonnes ; mais chacune porte une information spécifique. Or, faute d'avoir précisé laquelle de ces bases est choisie comme référence pour calculer les pourcentages, nombre de logiciels calculent les trois par défaut, alors qu'une seule base est sans doute pertinente pour étudier le problème posé.

Base « total des observations »

Cette base renseigne sur la structure globale de l'échantillon et mesure les disparités structurelles existantes (cellules sous- ou surreprésentées).

Dans l'échantillon étudié, les pourcentages entre parenthèses correspondent à cette base. Et on constate que 5,6 % des voyageurs du sexe féminin et 25,6 % des voyageurs du sexe masculin ont l'habitude de prendre un train de nuit : il y a donc dans l'échantillon 4,5 fois plus d'hommes que de femmes qui empruntent ce type de train.

Cette base peut aider à générer des hypothèses de travail (par exemple, s'interroger sur les raisons pour lesquelles les femmes fréquentent moins les trains de nuit) ou à sur-échantillonner les femmes pour obtenir une répartition équitable entre sexes pour étudier leurs attentes en matière de services spécifiques pour ce train.

Bases définies par les effectifs marginaux : colonnes et lignes

– *Sous le point de vue « colonne ».* Cette base de référence – où l'effectif d'une cellule est rapporté à l'ensemble des observations d'une modalité de la variable en colonne (39 pour la modalité « 4-nuit ») – permet d'isoler autant de sous-échantillons différents que de modalités contenues dans la variable (par exemple : les différents types de trains). Pour chacun de ces échantillons (définis par une modalité de la variable), on peut alors comparer la distribution des observations sur les modalités de l'autre variable.

Parmi les personnes qui ont l'habitude de prendre un « 4- train de nuit » (n = 39), on rencontre environ une seule femme (18 %) pour 4 hommes (82 %). Très logiquement, on retrouve le même ratio que précédemment : il y a 4,5 fois (82/18) plus d'hommes que de femmes pour emprunter ce type de train. Ce qui est environ deux fois plus que pour les trains du matin où on relève une femme (33 %) pour deux hommes (66 %), soit 4,5/2 = 2,25.

– *Sous le point de vue « ligne »* : l'effectif d'une cellule est rapporté à l'ensemble des observations qui sont caractérisées par une modalité de la variable en ligne. Cela permet de comparer entre chaque groupe défini par ces modalités (les hommes et les femmes dans l'exemple) le profil correspondant sur l'autre variable (les types de trains).

On constate que 40 % des hommes (32/80) ont l'habitude de prendre un train de nuit et les 60 % restant se partagent entre les trois autres types de train. De même, 42 % des femmes (19/45) ont plutôt l'habitude de prendre un train de fin de soirée et les 58 % restant se partagent les autres types de train. Ces deux types de train ont la même structure de distribution (environ 40 %-60 %), mais inversée selon le sexe considéré. Ce constat peut être la source d'hypothèses à formuler.

1.2 Disposer le tableau en cas de relation hiérarchique supposée

Décider quelle base prendre comme référence peut être la première difficulté à surmonter. En cas de relation symétrique entre les deux variables, cela ne change rien de raisonner sur les lignes ou sur les colonnes. Mais en cas de relation hiérarchique, il faut opter pour une présentation qui explicite bien le sens de cette hiérarchie.

© Dunod – Toute reproduction non autorisée est un délit.

Pratiquement, poser en ligne la variable indépendante (ou « explicative ») – celle dont dépendent les variations de l'autre variable – et en colonne la variable dépendante (ou « expliquée ») comme le présente le tableau 3.1, peut faciliter la lecture. Quoi qu'il en soit, chaque modalité de la variable explicative définit un sous-échantillon d'observations dont la distribution des fréquences, selon les modalités de l'autre variable, doivent être calculées sur une base 100 %. Ainsi, cela revient à comparer des distributions : la distribution de la fréquence des modalités de la variable « dépendante » (mise en colonne par exemple) est-elle comparable quelles que soient les modalités (en lignes) considérées de la variable « explicative » (et permettant de conclure à l'indépendance entre les deux variables), ou, au contraire, cette distribution varie-t-elle (laissant supposer une certaine dépendance) ?

Tableau 3.1 – Disposition du tableau croisé dans une logique causale

	« Habitude de choix d'un type de train »				
« Genre »	1- matin	2- journée	3- soirée	4- nuit	**Total**
1 - Féminin	10 22,2 % (8,0 %)	9 20,0 % (7,2 %)	19 42,2 % (15,2 %)	7 15,6 % (5,6 %)	**45** **100 %** **36,0 %**
2 - Masculin	20 25,0 % (16, 0 %)	12 15,0 % (9, 6 %)	16 20,0 % (12, 8 %)	32 40,0 % (25, 6 %)	**80** **100 %** **64,0 %**
Total	**30**	**21**	**35**	**39**	**125**

La première ligne de chaque cellule donne les effectifs observés dans l'échantillon (ex. : 10 femmes et 20 hommes prennent le train du matin) ; la deuxième ligne donne le pourcentage calculé sur le nombre d'observations en ligne (ex. : 45 = 100 %, d'où 10/45 = 22,2 %) et la troisième ligne donne le pourcentage entre parenthèses calculé sur le nombre total d'observations (ex. : 125 = 100 % d'où 20/125 = 16 %).

Dans l'exemple SNCF, la variable « genre » est posée en ligne et les pourcentages qui permettent de comparer la distribution du choix des trains pour chaque sexe sont calculés respectivement en fonction du nombre de femmes (n : 45 = 100 %) et du nombre d'hommes (n : 80 = 100 %). D'emblée, il est visible que la distribution du choix d'un type de train (variable en colonne) n'est pas identique selon les hommes et les femmes (variable en ligne). Cette lecture permet de constater que les femmes utilisent plutôt les trains de fin de journée et les hommes les trains de nuit ; ce constat légitime en quelque sorte l'hypothèse selon laquelle le genre aurait une incidence sur les habitudes de fréquentation d'un type de train spécifique. Mais cette hypothèse demande impérativement à être vérifiée à l'aide d'un test statistique : le test d'indépendance fondé sur le Chi-deux.

2 Le test d'indépendance fondé sur le Chi-deux (χ^2)

2.1 Comprendre la logique du test

Le test d'indépendance permet de vérifier s'il existe ou non une relation entre les *deux variables non métriques* qui structurent un tableau croisé. Notons que la

logique statistique infère que ces deux variables sont indépendantes, les variations de l'une n'entraînant pas de variations systématiques pour l'autre.

Le principe du test – qui s'inscrit dans la logique classique du test d'hypothèse statistique – consiste à mesurer les écarts entre une distribution observée à une distribution théorique (et non à comparer directement les distributions générées par les modalités d'une variable). Le Chi-deux est la statistique qui mesure ces écarts.

Pour un seuil de risque α, on peut donc formuler :

- une hypothèse nulle (H_0) selon laquelle les écarts entre les distributions, observée et théorique, ne sont pas significatifs d'une vraie différence, d'où, $H_0 : \chi^2 = 0$. Accepter cette hypothèse conduit à admettre l'indépendance des deux variables ;
- une hypothèse alternative (H_1) selon laquelle les écarts entre les distributions, observée et théorique, sont significativement différents de zéro et traduisent une réelle différence ; d'où $H_1 : \chi^2 \neq 0$. Accepter cette hypothèse conduit à admettre qu'il existe une relation significative entre les deux variables étudiées.

Le χ^2 ayant sa propre loi de distribution, on peut donc décider, au vu de sa valeur calculée pour les données traitées, laquelle des deux hypothèses, nulle ou alternative, est la plus plausible.

La distribution théorique des observations entre les k cellules résultant du croisement des deux variables correspond à l'effectif calculé à partir de la probabilité qu'a un élément de la population de correspondre aux modalités des deux variables, compte tenu de l'effectif total. L'effectif théorique des modalités est calculé à partir des effectifs marginaux (effectifs en ligne et en colonne) et du nombre total d'observations (voir la rubrique « Focus »).

Focus

Les différentes étapes du calcul du test d'indépendance

Bases du test d'indépendance :

Le calcul du χ^2 est basé toujours sur les effectifs : un tableau de données fournis en % doit donc impérativement être transformé en un tableau d'effectifs (ce qui suppose de connaître au moins l'effectif total à la base du calcul de ces pourcentages). Le principe du calcul pour une cellule est simple :

(Effectifs Observés – Effectifs Théorique)²/Effectifs Théoriques

Appliquée à l'ensemble des cellules, le test du Chi-deux (χ^2) est formulé comme suit :

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(n_{ij}^{obs} - n_{ij}^{th}\right)^2}{n_{ij}^{th}}$$

☞

© Dunod – Toute reproduction non autorisée est un délit.

où : n_{ij}^{obs} - effectifs observés ; n_{ij}^{th} effectifs théoriques ; p et q sont le nombre des catégories des variables X et Y respectivement. Chaque cellule du tableau de contingence contient les observations qui partagent ensembles une catégorie de chaque variable.

Le Chi-deux est sensible au nombre d'observations et sa fiabilité impose au moins 5 observations par cellule (ou que 20 % des cellules n'aient pas un effectif inférieur à 5 dans le cas de tableaux croisant deux variables aux nombreuses modalités).

Le nombre total de cellules est le résultat du produit entre le nombre de catégories en ligne (l) et en colonne (c). Ce nombre permet de calculer les degrés de liberté nécessaires pour déterminer la valeur critique théorique ; le degré de liberté (abrégé en « ddl » - ou « df » en anglais, pour « degree of freedom ») se calcule simplement : (l -1) * (c -1), soit dans l'exemple traité : (2 -1) * (4 – 1) = 3.

Calcul de l'effectif théorique des cellules k (= distribution théorique) :

Formule : (Total effectifs en Ligne x Total effectifs en Colonne)/Effectif Total

Appliqué au cas SNCF, voici ce que cela donne pour les 8 (=2*4) cellules :

1 x 1 : (45 * 30) / 125 = 10,80	2 x 1 : (80 * 30) / 125 = 19,20
1 x 2 : (45 * 21) / 125 = 7,56	2 x 2 : (80 * 21) / 125 = 13,44
1 x 3 : (45 * 35) / 125 = 12,60	2 x 3 : (80 * 35) / 125 = 22,40
1 x 4 : (45 * 39) / 125 = 14,04	2 x 4 : (80 * 39) / 125 = 24,96

Calcul du Chi-deux :

Formule : *Somme (Effectifs Observés – Effectifs Théorique)² / Effectifs Théoriques*

Le χ^2 total est la somme des χ^2 partiels calculés pour chacune des cellules du tableau :

χ^2 Variable « sexe » / féminin :		χ^2 Variable « sexe » / masculin :	
1 x 1 : (10 – 10,80)² / 10,80 = 0,06	+	2 x 1 : (20 – 19,20)² / 19,20 = 0,03	+
1 x 2 : (9 – 7,56)² / 7,56 = 0,27	+	2 x 2 : (12 – 13,44)² / 13,44 = 0,15	+
1 x 3 : (19 – 12,60)² / 12,60 = 3,25	+	2 x 3 : (16 – 22,40)² / 22,40 = 1,83	+
1 x 4 : (7 – 14,04)² / 14,04 = 3,53	=	2 x 4 : (32 – 24,96)² / 24,96 = 1,99	=
Total « féminin » = 7,11	+	Total « masculin » = 4,01	=
χ^2 total = 0,06 + 0,27 + 3,25 + 3,53 + 0,03 + 0,15 + 1,83 + 1,99 = 11,12			

Détermination du seuil de risque α, des degrés de liberté, de la valeur critique du χ^2 et interprétation du résultat statistique :

- seuil de risque (probabilité de commettre une erreur en acceptant H_0 quand cette hypothèse est fausse) habituellement fixé à 5 % ($\alpha = 0,05$) ;
- degrés de liberté liés au nombre de cellules du tableau : *(2 -1) * (4 – 1) = 3*
- Valeur critique théorique du χ^2 à $\alpha = 0,05$ et 3 d.d.l., lue dans la table correspondante : 7,82
- La valeur du χ^2 calculée (11,12) est supérieure à la valeur théorique (7,82) : l'hypothèse Ho d'indépendance entre les deux variables n'est pas vérifiée, rendant H1 plausible : les deux variables sont significativement liées.

☞

Le calcul du χ^2 dans le cas particulier de deux variables binaires générant un tableau carré 2x2.

Certaines variables sont binaires (nécessitant des réponses de type « oui/non », « d'accord/pas d'accord », « possède/ne possède pas » une caractéristique, etc.) et on souhaite déterminer leur degré d'association. L'exploitation des tableaux croisés symétriques (appelés aussi « tableau carré » puisque le nombre de modalités est le même selon les lignes et les colonnes) peut se faire avec des outils spécifiques.

Au test d'indépendance classique fondé sur le χ^2, on pourra apporter la correction de Yates, particulièrement lorsque certaines cellules ont un petit effectif (inférieur à 10 unités, à l'exemple du cas SNCF) ; cette correction consiste à retrancher la quantité 0,5 à l'écart absolu au sein d'une cellule :

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(\left|n_{ij}^{obs} - n_{ij}^{th}\right| - 0,5\right)^2}{n_{ij}^{th}}$$

Certains logiciels proposent une autre procédure avec l'application du test exact de Fisher spécifique aux tableaux 2 x 2 ; il précise la probabilité critique unilatérale.

Tableau 3.2 – Exemple de tableau carré issu du croisement entre « genre » et « type de train » pour les seuls trains de soirée et de nuit

	train soirée	train nuit	Total en ligne	
	A	B		Contenu des lignes :
1 - féminin	19	7	26	effectif observé
	73,08 %	26,92 %	35,14 %	% en ligne
	12,30	13,70		effectif théorique
	3,65	3,28		Chi-deux partiel
	C	D		
2 - masculin	16	32	48	effectif observé
	33,33 %	66,67 %	64,86 %	% en ligne
	22,70	25,30		effectif théorique
	1,98	1,78		Chi-deux partiel
Total en colonne	35	39	74	
	47,30 %	52,70 %	100,00 %	

Résultat des tests

Test du Chi-deux :

Chi-deux	d.d.l.	Proba.
10,69	1	0,0011 (sans correction de Yates)
9,15	1	0,0025 (avec correction de Yates)

Test exact de Fisher pour des tableaux 2x2 : probabilité P unilatérale = 0.0011

Conclusion : au seuil de risque $\alpha = 0,05$, on rejette l'hypothèse d'indépendance pour accepter l'hypothèse d'une relation significative entre les deux variables.

© Dunod – Toute reproduction non autorisée est un délit.

2.2 Les résultats du test

La décision statistique

La décision d'accepter l'hypothèse nulle (H_0 : indépendance) ou l'hypothèse alternative (H_1 : relation) dépend de la valeur du χ^2 calculé. On accepte l'hypothèse H_0 de la nullité des écarts entre les deux distributions observée et théorique lorsque pour un seuil de risque déterminé :

– soit la *valeur critique* calculée du χ^2 est inférieure à sa valeur critique théorique (déterminée par lecture de la table correspondante) ; si $\chi^2_{cal} \leq \chi^2_{th}$, alors H_0 est acceptée : les deux variables sont indépendantes, sans lien ;
– soit la *probabilité critique* (p.c.) de la valeur calculée du χ^2 est supérieure au seuil de risque fixé α ; puisque p.c. $\geq \alpha$, alors on rejette H_0 au bénéfice de H_1 : les deux variables sont liées et dépendantes l'une de l'autre.

– **Dans l'exemple traité**, pour un seuil de risque usuel de 5 % ($\alpha = 0{,}05$) et pour 3 degrés de liberté, on se doit de rejeter H_0 au bénéfice de H_1 et admettre que les deux variables sont liées ; les deux modes conduisent à la même décision statistique :
– selon la valeur critique calculée du Chi-deux : la valeur calculée (χ^2_{cal} : 11,12) est supérieure à la valeur théorique (χ^2_{th} : 7,82) (lisible dans la table du χ^2 - voir Annexes) ;
– selon la probabilité critique calculée, cette valeur (fournie par le logiciel utilisé pour calculer le test) valant 1,11 % (χ^2_{cal} : 0,0111) est inférieure au seuil de risque de 5 % retenu. Un seuil de risque plus sévère, fixé à 1 % ($\alpha = 0{,}01$) conduirait à conclure de manière différente et admettre l'indépendance, car 1,11 % est supérieur à 1,0 %.

L'interprétation du résultat du test

Si le résultat statistique fondé sur le test du Chi-deux permet de décider qu'il existe un lien significatif entre les deux variables non-métriques étudiées, il ne renseigne cependant pas sur l'intensité de la relation (c'est une limite de ce test), ni ne précise comment les résultats doivent être interprétés. Or, l'intérêt d'une telle démarche est de permettre à l'analyste de porter des conclusions opérationnelles (pour susciter une réflexion ou orienter des décisions).

En s'appuyant sur l'exemple du choix d'un type de train, nous décrivons une méthode qui permet de porter des conclusions raisonnables à partir d'une exploitation systématique du tableau de résultat des valeurs du χ^2.

La valeur du Chi-deux résulte de la somme de chacun des écarts constatés dans chaque cellule du tableau croisé entre les effectifs observés et théoriques (*cf.* tableau 3.1). On dispose ainsi, cellule par cellule, d'une valeur partielle de ce χ^2 appelée « Chi-deux partiel ». Le ratio de chaque χ^2 partiel au χ^2 total mesure la contribution d'une cellule aux écarts constatés entre les distributions observée et théorique, et c'est sur ce ratio que repose le travail. La méthode consiste donc à :

1. identifier les cellules dont les contributions sont les plus élevées (ces ratios n'ont qu'une valeur indicative et comparative) ;

2. identifier dans ces cellules le signe des écarts résultant de la différence calculée entre les effectifs observés et les effectifs théoriques ; le signe « + » indique une surreprésentation des observations : il y en a plus que, théoriquement, il ne devrait ; inversement, le signe « - » indique une sous-représentation des observations : il y en a moins que, théoriquement, il ne devrait dans la cellule.

À considérer les valeurs du Chi-deux pour chaque cellule, telles que rapportées dans le tableau 3.3, on constate :

– des écarts limités pour quatre cellules (χ^2 partiels compris entre 0,03 et 0,27 avec des ratios de contribution inférieurs à 2,5 %) : celles relatives aux trains du matin et de mi-journée ; la distribution des habitudes de fréquentation de ces trains ne varie pas en fonction du sexe du voyageur, les habitudes sont pratiquement les mêmes.

– des écarts conséquents pour les quatre autres cellules (χ^2 partiels variant de 1,83 à 3,53 ; ratios supérieurs à 16,5 %) qui concernent les trains fréquentés en soirée et de nuit ; ces écarts sont encore plus importants chez les femmes que chez les hommes ; la distribution des habitudes de fréquentation de ces trains varie sensiblement en fonction du sexe.

Tableau 3.3 – Données pour l'interprétation des résultats avec les Chi-deux partiels

« Genre »	Informations :	« Habitudes d'un type de train »			
		1- matin	2- journée	3- soirée	4- nuit
1- Féminin	Effectif observé	10	9	19	7
	Effectif théorique	10,80	7,56	12,60	14,04
	Signe de l'écart	-	+	+	-
	Chi-deux partiel	**0,06**	**0,27**	**3,25**	**3,53**
	Ratio χ^2 partiel *	0,6 %	2,5 %	29,2 %	31,7 %
2- Masculin	Effectif observé	20	12	16	32
	Effectif théorique	19,20	14,44	22,40	24,96
	Signe de l'écart	+	-	-	+
	Chi-deux partiel	**0,03**	**0,15**	**1,83**	**1,99**
	Ratio χ^2 partiel	0,3 %	1,3 %	16,5 %	17,9 %
* ratio = χ^2 partiel / χ^2 total		χ^2 total = 11,12 soit la somme des χ^2 partiels			

Puisqu'il y a un lien statistiquement significatif entre le choix d'un type de train et le genre du voyageur (valeur globale du χ^2), on peut donc conclure, en considérant simultanément la valeur absolue du χ^2 partiel et le signe de l'écart entre les effectifs observés et théoriques, que :

a) les différences d'habitude qui existent entre les hommes et les femmes ne concernent que les trains de seconde partie de journée (soirée et nuit) : χ^2 partiels les plus élevés ;

© Dunod – Toute reproduction non autorisée est un délit.

b) les femmes ont moins l'habitude de fréquenter les trains de nuit qu'elles ne devraient « théoriquement » le faire (– | 3,53 |), au bénéfice des trains de fin de journée (+ | 3,25 |) ; les hommes ont plus l'habitude de fréquenter les trains de nuit qu'ils ne devraient « théoriquement » le faire (+ | 1,9 |) au détriment des trains de fin de journée (– | 1,88 |) ;
c) ces différences d'habitudes comportementales concernent plus encore les femmes que les hommes : la contribution des écarts au total pour les femmes est presque deux fois supérieure (6,78/3,82 = 1,77) à celle des hommes.

Il apparaît que le sexe est une variable pertinente pour segmenter les habitudes de fréquentation, mais son « effet » est limité aux seuls trains de seconde partie de journée : en effet, les femmes fréquentent significativement plus les trains de début de soirée que les trains de nuit (42,2 % d'entre elles) quand les hommes ont les pratiques inverses (40 % d'entre eux ont l'habitude des trains de nuit) ; mais il n'y a pas de différence significative dans les habitudes de fréquentation des autres types trains (du matin et de mi-journée).

Partant de ce constat, il est donc légitime pour la direction commerciale de la SNCF de s'inquiéter plus spécifiquement auprès de la clientèle féminine des raisons d'une moindre fréquentation des trains de nuit et d'évaluer quels services ou conditions permettraient de les y amener.

3 Les autres tests d'association entre deux variables qualitatives

3.1 La mesure de l'association entre deux variables non métriques

Deux indicateurs ont été construits sur la base du Chi-deux pour mesurer l'intensité de la liaison entre les deux variables associées (ce que le test d'indépendance ne peut faire) : le coefficient de contingence C et le coefficient V de Cramer (présentés dans le tableau 3.4) dont le calcul est effectué par les logiciels.

Tableau 3.4 – Indices de mesure de l'association entre deux variables fondés sur le Chi-deux

Le coefficient C de contingence	Le coefficient V de Cramer
$C = [\chi^2 / (\chi^2 + n)]^{1/2}$ n : la taille de l'échantillon.	$V = (\chi^2 / nt)^{1/2}$ n : la taille de l'échantillon et t : valeur minimum des degrés de liberté liés aux lignes et colonnes ; prendre la plus petite valeur de (l – 1) ou (c – 1)
Interprétation : C = 0 quand $\chi^2 = 0$: absence d'association Mais sa limite supérieure est inférieure à 1 et dépend de la variable ayant le plus petit nombre de modalités k ; elle se calcule ainsi : $[(k - 1) / k]^{1/2}$	Interprétation : Coefficient normé variant entre 0 et 1 quelle que soit la taille du tableau : V = 0 : absence d'association V = 1 : association maximale

Pour l'exemple traité dans cette première section, le coefficient de contingence (ou coefficient d'incertitude) vaut 0,258 ; et sa limite supérieure vaut 0,707, soit $[(2 - 1) / 2]^{1/2}$; aussi ce 0,258 doit être interprété comme traduisant une faible association, résultat confirmé par le coefficient de Cramer qui vaut 0,298 (sur un intervalle variant entre 0 et 1). L'examen des écarts entre effectifs observés et théoriques dans chaque cellule (tableau 3.3) a montré que les liens significatifs ne concernent que les trains de seconde partie de journée. Un test d'indépendance effectué sur le tableau carré n'intégrant que ces deux seules modalités donne un résultat hautement significatif : $\chi^2_{cal} = 10{,}69$ ($> \chi^2_{th} = 6{,}64$ pour $\alpha = 1$ %) avec un seul degré de liberté (la valeur du χ^2 calculé correspond à une probabilité critique de l'ordre de 0,1 % [0,0011]) ; la valeur des deux coefficients d'association s'améliorent de près de 30 % (respectivement : C = 0,335 et V = 0,380).

3.2 L'utilisation des informations dans une perspective prédictive : le coefficient Lambda (Λ) de Guttman

Lorsque deux variables sont significativement liées comme un test d'indépendance peut le démontrer, l'information disponible sur l'une d'elle renseigne pour partie sur l'autre. Dans la situation extrême d'une association parfaite (le coefficient de Cramer vaut 1), on pourrait donc prédire les valeurs des modalités de la seconde variable connaissant celles de la première. Sachant par exemple que, dans le choix d'un type de train, il existe un lien entre le choix fait pour l'aller et celui pour le retour, connaissant alors le train emprunté à l'aller, on devrait pouvoir anticiper celui utilisé pour le retour. Cette perspective prédictive fondée sur l'étude de variables non métriques peut être pertinente dans diverses situations.

Pour répondre à cette préoccupation, il existe un indicateur d'association prédictive, le *coefficient Lambda* (Λ) de Guttman qui mesure l'utilité d'une variable pour prédire l'autre ; de manière concrète, il indique dans quelle mesure une ligne du tableau croisé est utile pour prévoir les valeurs de la variable en colonne, et inversement. S'agissant d'un indicateur dissymétrique, les résultats ne sont pas les mêmes selon que la ligne ou la colonne est considérée comme étant la variable prédictive. Gardant toujours le même exemple, le tableau 3.5 donne les valeurs de ce coefficient pour le tableau croisé complet (A : 2 x 4) et pour le tableau carré n'intégrant que les trains de soirée et de nuit (B : 2 x 2).

Tableau 3.5 – Indices de mesure de l'association entre deux variables fondés sur le Chi-deux

A - Tableau croisé complet (2 x 4)				B - Tableau croisé réduit (2 x 2)			
	symétrique	avec lignes dépendantes	avec colonnes dépendantes		symétrique	avec lignes dépendantes	avec colonnes dépendantes
Lambda	0,1145	0,0667	0,1395	lambda	0,2459	0,1154	0,3429

© Dunod – Toute reproduction non autorisée est un délit.

Dans les deux cas de figure, on constate que la variable en ligne donne une meilleure prédiction des résultats de la variable en colonne (« dépendante ») que l'inverse, particulièrement quand l'étude est limitée aux seuls trains de soirée et de nuit. Ces résultats confortent aussi l'hypothèse selon laquelle l'habitude d'usage d'un type de train est bien liée au sexe du voyageur et a une incidence sur les choix effectués.

Une utilisation dérivée de ce coefficient Lambda d'association prédictive est donc de vérifier le sens de la hiérarchie et la logique de la dépendance entre les deux variables étudiées simultanément. Ici, c'est bien le choix du type de train (variable dépendante en colonne) qui varie selon le sexe du voyageur (variable indépendante en ligne).

3.3 La mesure de l'association entre deux variables binaires

Le *coefficient Q de Yule* permet de mesurer l'association entre deux variables binaires. Il varie entre 0 (pas d'association) et 1 (association maximale) et son calcul est simple (*cf.* tableau 3.2) :

Soit les cadrans du tableau carré désignés par les lettres A B C D, alors le coefficient Q se calcule selon la formule :

$$(AD - BC) / (AD + BC)$$

> Dans l'exemple suivi : $Q = [(19*32) - (7*16)] / [(19*32) + (7*16)] = 0{,}689$, ce qui témoigne d'une forte association entre les deux variables « sexe » et « type de train » (0,689), en cohérence avec les résultats du test d'indépendance, sans que l'on puisse cependant éprouver sa significativité (comme permettrait de le faire un coefficient de corrélation).

Le test d'indépendance fondé principalement sur le Chi-deux est bien adapté pour traiter des tableaux croisés de petite taille, à l'exemple du cas exploité. Mais lorsque chacune des deux variables comporte un grand nombre de modalités, un autre outil – l'*analyse des correspondances* – permet d'analyser globalement et avec efficacité, les diverses associations possibles entre les différentes modalités des deux variables. C'est l'objet de la section suivante.

L'ESSENTIEL

■ ***Les tests pour analyser la relation entre deux variables non métriques***

– Il est possible de tester l'existence d'une relation entre paires de variables non-métriques à l'aide d'un test d'indépendance, tel que le Chi-deux (χ^2) appliqué sur les données brutes (effectifs) d'un tableau croisé.

– Dans la mesure où la précaution a été prise de construire un tableau croisé en identifiant la variable qui est susceptible de rendre compte des variations de l'autre variable (logique causale), ce test assez simple permet de supputer l'hypothèse d'une relation de causalité (sans toutefois la démontrer). Dans cette logique, le coefficient Lambda (Λ) permet d'évaluer la performance prédictive de la variable indépendante ou « explicative ».

– Le test du Chi-deux ne peut pas mesurer l'intensité de l'association entre deux variables non-métrique (à l'instar du coefficient de corrélation pour deux variables métriques), mais d'autres tests dérivés du χ^2 permettent de le faire (coefficients de contingence et coefficient de Cramer).

Section 2 L'ANALYSE DES CORRESPONDANCES POUR EXPLOITER UN TABLEAU DE CONTINGENCE

Étudier la relation entre deux variables non-métriques avec un test d'indépendance est assez aisé si le nombre de modalités reste limité ; mais il n'en est plus de même si ce nombre augmente. Or, pour comparer des groupes d'individus (des cibles) ou des objets (des produits, des marques) sur la base de caractéristiques ou d'attributs qui les décrivent, les variables utilisées peuvent comporter de nombreuses modalités.

L'analyse des correspondances simples, qui repose sur la métrique du Chi-deux, est l'outil adapté dans un tel contexte et il répond à trois objectifs :

– tester la significativité de l'association entre les deux variables non-métriques ;
– étudier les liens entre les modalités des deux variables, et aussi au sein de chaque variable ;
– visualiser graphiquement les liens significatifs entre ces modalités à l'aide d'une carte perceptuelle – un « mapping » – dont l'objectif est de rendre l'interprétation des résultats plus aisée (malgré les risques associés à une lecture trop rapide).

© Dunod – Toute reproduction non autorisée est un délit.

En effet, la pertinence des interprétations qui découlent de ces cartes – et donc les décisions managériales qui s'en suivent – dépend du soin apporté à leur réalisation. Aussi, après avoir présenté la logique du modèle et le type de résultats fournis, nous proposerons une approche pour la lecture et l'interprétation rigoureuse des résultats visuels obtenus en s'appuyant sur le traitement d'un cas concret.

Cas d'illustration

La perception des plats cuisinés selon leur mode de conservation

Pour connaître le positionnement distinctif des plats cuisinés consommés dans les foyers, une étude a été conduite auprès d'un échantillon de 245 individus. Les personnes enquêtées devaient indiquer dans quelle mesure chacun des 11 attributs utilisés correspondait ou non à chacun des six types de conserve soumis à leur jugement. Les résultats sont présentés dans le tableau ci-dessous :

« convient bien avec » : (tableau des effectifs)	Conserve « maison »	Conserve appertisée	Plat traiteur	Plat surgelé	Plat déshydraté	Plat frais sous vide
1 Qui permet de se faire **plaisir**	223	91	211	159	47	83
2 Qui offre un bon rapport **qualité/prix**	201	145	54	137	59	71
3 Qui permet de déguster des **plats originaux**	201	93	186	176	39	76
4 Qui offre des plats légers de **bonne qualité**	149	39	69	181	22	76
5 Plats de qualité, **savoureux**	225	74	181	145	27	83
6 Qui se conserve longtemps, qu'on **stocke**	37	218	12	189	162	44
7 **Pratique** si pas le temps de préparer un repas	20	221	154	216	162	142
8 Plats qui sont **rapidement prêts**	27	228	179	186	189	169
9 Qui offre garantie de fraîcheur, de **naturel**	221	59	120	167	22	86
10 Qui offre un grand choix de **recettes**	203	108	132	169	39	66
11 Qui permet de satisfaire **tous les goûts**	221	127	145	172	49	83

Avec cette étude, l'industriel commanditaire souhaitait connaître le système de représentation que les consommateurs avaient des six modes de conservation pour bien situer différentes offres de plats surgelés et orienter en conséquence les stratégies de positionnement de sa marque.

1 Les bases de l'analyse factorielle des correspondances simples

1.1 Réorganiser les données pour obtenir un tableau de contingence

L'*analyse factorielle des correspondances simples* (symbolisée AFC et parfois appelée analyse factorielle binaire) traite un tableau de contingence croisant deux variables non métriques (ou rendues telles par discrétisation). Les modalités de l'une des variables correspondent aux observations (des individus, des marques ou des produits) et celles de l'autre, aux caractéristiques utilisées pour les décrire. Chaque cellule du tableau croisé contient les effectifs issus du comptage : soit le nombre de fois où une caractéristique a été associée à une observation, un item à un produit par exemple. L'objectif du traitement est d'identifier les attributs qui caractérisent le mieux chacun des produits comparés.

> Dans l'exemple des plats cuisinés, il s'agit de voir comment un mode de fabrication d'un plat est associé aux différents items. Ainsi, 223 répondants pensent que l'attribut « permet de se faire plaisir » convient bien avec la « conserve maison » ; réciproquement, 22 personnes (= 245-223) pensent que cet item ne convient pas à ce type de conserve ; de même, 83 personnes pensent que les « plats frais sous vide » permettent de « satisfaire tous les goûts » et donc 162 (= 245-83) pensent qu'ils ne le permettent pas.

Dans ce tableau de contingence, chacune des deux variables a le même statut : il n'y a pas de hiérarchie établie où les variations de l'une pourraient expliquer celles de l'autre. Elles correspondent chacune à un contenu d'information spécifique :
– pour la variable « objet » : chaque observation, marque ou produit, définit une modalité ;
– pour la variable « caractères » : chaque attribut retenu constitue une modalité.

> **Exemple.** Dans l'étude « plats cuisinés », chacun des 11 items constituait une question indépendante, avec une échelle de réponse ordinale en 5 points allant de « convient très bien » à « ne convient pas du tout ». Pour constituer le tableau de contingence, n'ont été retenues que les réponses exprimant un accord (« convient très bien » et « convient bien ») pour chacun des items. Ainsi, par construction, chacun des onze items devient une des modalités d'une variable dénommée « variable d'image ». C'est un exemple illustrant la construction d'une variable tel qu'abordé dans le chapitre 1.

1.2 La logique du modèle

Dans l'hypothèse où le tableau de contingence décrit des objets (modalités d'une variable) par un ensemble de caractères (modalités de l'autre variable) :
– *deux objets sont proches* si la distribution de leurs effectifs sur l'ensemble des caractères (leur profil) est presque identique ; ils se confondent si leurs deux profils de distribution sont rigoureusement identiques (distance nulle).

> **Exemple :** « plat déshydraté » et « plat frais sous vide » ont des profils voisins en comparaison des autres produits.

© Dunod – Toute reproduction non autorisée est un délit.

– *deux objets sont éloignés*, voire opposés, si la distribution de leurs effectifs sur les caractères est très différente, voire systématiquement opposée :

> **Exemple :** « plat surgelé » et « plat frais sous vide » ont un profil assez différent qui les différencie nettement.

Le même type d'analyse peut être conduit avec les caractères : deux items sont proches, ou éloignés, selon que le profil de distribution de leurs effectifs selon les différents objets est comparable ou différent.

> **Exemple :** on peut constater une certaine proximité entre les attributs « 1-plaisir » et « 5-savoureux » et une opposition entre les attributs « 8-rapidité » et « 9-naturel ».

L'analyse des correspondances vise à rendre visible ces ressemblances ou ces dissemblances en mesurant la distance qui existe entre chaque profil de distribution des modalités en ligne d'une part, des modalités en colonne d'autre part, puis entre les modalités associant lignes et colonnes simultanément. Comme déjà vu dans la section précédente, le Chi-deux (χ^2) permet de mesurer la distance qui existe entre deux profils de distribution : si deux profils se confondent, cette distance est nulle ($\chi^2 = 0$) ; et sa valeur croit au fur et à mesure que les deux profils comparés se différencient ; elle atteint son maximum quand ils sont symétriquement opposés. C'est précisément cette matrice de distances entre les modalités fondée sur le Chi-deux qui est à la base du modèle mathématique de l'AFC (*cf.* Pour aller plus loin).

L'analyse factorielle des correspondances consiste donc d'abord à calculer toutes les distances (sur la base du χ^2) entre l'ensemble des profils « ligne » et « colonne » générés par la structure du tableau de données et à effectuer ensuite sur ces nouvelles données de nature métrique (puisque des distances) deux analyses en composantes principales (voir chapitre 4) : une première sur les modalités en ligne (« individus ») où les modalités en colonne jouent le rôle des variables ; une seconde qui inverse les rôles où les modalités en ligne prennent valeur de variable et celles en colonne, d'individus.

Pour aller plus loin – Les principes mathématiques de l'Analyse factorielle des correspondances simples (AFC)

L'analyse des correspondances n'est qu'un cas particulier de l'analyse en composantes principales (*cf.* chapitre 4) qui traite des données quantitatives. La première opération consiste donc à transformer chaque caractère initialement qualitatif en autant de caractères quantitatifs qu'il y a de catégories pour chacune des deux variables. C'est donc le tableau de contingence mettant en ligne les p catégories de la variable X et en colonne les q catégories de la variable Y qui constitue la base initiale de calcul de l'AFC. En réalité l'AFC travaille sur le tableau des probabilités (voir tableau ci-dessous) qui est obtenu en divisant toutes les cellules du tableau de contingence par l'effectif total n (nombre des unités statistiques étudiées). On obtient alors les probabilités conjointes $f_{ij} = n_{ij}/n$ de posséder à la fois la catégorie i de la première variable et la

☞

catégorie j de la seconde variable. La somme de tous les f_{ij} est alors 1. La marge en colonne de ce tableau notée f_i représente la somme des termes de lignes alors que la marge ligne notée f_j représente la somme des termes des colonnes.

		Catégories de la variable Y					Marge
		1	…	j	…	q	
Catégories de la variable X	1	f_{11}	…	f_{1j}	…	f_{1q}	$f_{1.}$
	…	…	…	…	…	…	…
	i	f_{i1}	…	f_{ij}	…	f_{iq}	$f_{i.}$
	…	…	…	…	…	…	…
	p	f_{p1}	…	f_{pj}	…	f_{pq}	$f_{p.}$
Marge		$f_{.1}$	…	$f_{.j}$	…	$f_{.q}$	1

L'étape suivante consiste à construire les profils ligne et les profils colonne. Les profils ligne sont obtenus en divisant chaque terme d'une ligne par sa marge $f_{ij}/f_{i.}$. On obtient ainsi des distributions de probabilités conditionnelles, la somme des probabilités étant égale à 1. Ainsi, la ligne *i* représente la distribution conditionnelle de la variable Y, sachant que l'on possède la catégorie *i* de la variable *X*.

De manière similaire, les profils colonne sont obtenus en divisant chaque terme d'une colonne par sa marge $f_{ij}/f_{.j}$. Les profils ligne peuvent alors être comparés avec le profil ligne moyen qui n'est rien d'autre que la distribution marginale de la seconde variable. De même, les profils colonne peuvent être comparés avec le profil colonne moyen qui représente la distribution marginale de la première variable. En comparant les profils ligne (colonne) avec le profil moyen ligne (colonne) l'AFC procède ainsi à une analyse multidimensionnelle de l'écart à l'indépendance.

L'AFC consiste alors à effectuer deux analyses factorielles en composantes principales : une sur les profils lignes et une sur les profils colonne. Il y a cependant deux différences notables entre l'AFC et l'ACP. La première porte sur le calcul de l'inertie des données. L'écart à l'indépendance d'un tableau de contingence est le Chi-deux :

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

Sachant que $f_{ij} = n_{ij}/n$ (donc $n_{ij} = f_{ij}/n$) et que les effectifs sous l'hypothèse d'indépendance sont $n_{ij}^{th} = f_{.j}f_{i.}$, on transforme l'expression ci-dessus sous la forme :

$$\chi^2 = n\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(f_{ij} - f_{.j}f_{i.}\right)^2}{f_{.j}f_{i.}} = n\,\phi^2$$

C'est justement le coefficient $\phi^2=\chi^2/n$, qui mesure intensité de liaison (ou l'intensité de l'écart à l'indépendance) entre les deux variables catégorielles X et Y – indépendant de la taille de l'échantillon puisque divisé par n – qui donne l'inertie du nuage des points des profils ligne et colonne.

L'autre spécificité de l'AFC par rapport à l'ACP et la mesure de distance entre les profils ligne (colonne) et le profil moyen ligne (colonne). L'AFC utilise la distance du Chi-deux à la place de la distance euclidienne. La distance du Chi-deux entre le profil ligne *i* et le profil ligne *i'* est donnée par la relation :

© Dunod – Toute reproduction non autorisée est un délit.

☞

$$d(i,i') = \sum_{j=1}^{q} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} + \frac{f_{i'j}}{f_{i'.}} \right)^2$$

De manière similaire, la distance du Chi-deux entre le profil colonne j et le profil ligne j' est donnée par la relation :

$$d(j,j') = \sum_{j=1}^{q} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} + \frac{f_{ij'}}{f_{.j'}} \right)^2$$

Une propriété très importante de l'AFC qui la distingue des autres analyses factorielles et que les profils ligne et les profils colonne jouent un rôle symétrique et que l'inertie des points ligne est la même que celle des points colonnes. En d'autres termes, les deux tableaux (profils ligne et colonne) ont exactement les mêmes valeurs propres. L'AFC décompose alors cette inertie par analyse factorielle en projetant le nuage des points des profils ligne (ou colonne) sur une suite d'axes orthogonaux dont l'origine est le profil moyen ligne (ou colonne). Il est possible de représenter simultanément les catégories des deux variables catégorielles sur le même plan factoriel, ce qui constitue la spécificité de l'AFC.

2 Les différentes étapes pour analyser les résultats d'une AFC

Quel que soit l'algorithme utilisé, les calculs fournissent trois types de résultats dont l'exploitation progressive va permettre de construire, d'interpréter et de valider un modèle statistique susceptible de rendre compte de manière économique de la plus grande part de la variance contenue dans le tableau de contingence étudié :

– Le **tableau des *valeurs propres*** nécessaire pour déterminer le nombre d'axes[1] significatifs et pour évaluer la qualité statistique du modèle construit ;
– Trois matrices contenant l'information nécessaire pour analyser les relations des axes avec toutes les modalités des deux variables croisées, puis interpréter les résultats avec rigueur, à savoir :

1. la **matrice des « *contributions absolues*** » (symbolisées CTR), pour identifier les modalités qui contribuent significativement à l'émergence d'un axe, contributions aussi appelées, vu leur rôle : « points explicatifs » ;
2. la **matrice des « *contributions relatives*** » ou cosinus-carrés (COS²) pour apprécier dans quelle mesure chacun des axes prend en charge une partie de la variance propre à chaque modalité ;
3. la **matrice des *coordonnées*** de chaque modalité sur chacun des axes qui permettent de construire les cartes perceptuelles.

1. Dans la littérature, les termes *axe* (utilisé dans ce chapitre), *dimension*, *facteur* et *composante* sont synonymes.

– Enfin, des graphiques ou ***cartes perceptuelles*** (ou « mapping ») qui visualisent la projection des modalités des deux variables dans des plans croisant les axes pris deux à deux.

Ces classes d'information définissent en fait les quatre étapes à suivre pour une analyse et une interprétation rigoureuses des résultats obtenus.

2.1 Première étape : construire et valider le modèle de représentation

La matrice des valeurs propres sert d'abord à identifier les axes significatifs et, ensuite, à évaluer la qualité du modèle retenu.

La signification des valeurs propres

La « trace » – ou « inertie » – mesure le degré de liaison entre les deux variables qualitatives étudiées : elle est nulle en cas d'indépendance totale entre les deux variables ; en cas de liaison, elle vaut χ^2 / n, où n est l'effectif total du tableau croisé et χ^2 la valeur du test d'indépendance.

> Exemple des « plats cuisinés » : Effectif n = 8 267 opinions exprimées et χ^2 = 1 577,09 : la trace ou inertie vaut[1] donc 1 577,09/8 267 = 0,19084.

L'inertie correspond en quelque sorte à la variance totale contenue dans le tableau de données initial (soit 0,19084 dans l'exemple : *cf.* tableau 3.6). L'inertie calculée se répartit entre plusieurs axes (lesquels correspondent aux vecteurs propres d'une matrice symétrique).

La *première colonne* du tableau 3.6 indique le nombre d'axes nécessaires pour rendre compte de la totalité de l'inertie. Par construction, le nombre total d'axes d'un tableau croisé correspond au nombre de modalités de la variable qui en a le moins auquel on retranche une unité (un degré de liberté) :

> Dans l'exemple, la plus « petite » des deux variables (type de conservation) n'ayant que 6 modalités, 5 axes (= 6-1) sont donc tout au plus suffisants pour rendre compte de la totalité de l'inertie, des 100 % de la variance qui vaut 0,19084.

Par construction, ces axes sont tous orthogonaux entre eux : ils sont donc indépendants et chacun apporte une information différente, complémentaire à celle déjà portée par les axes précédents : l'axe 2 prend en charge une partie de la variance résiduelle dont l'axe 1 ne rend pas compte.

1. Cette formule permet donc de calculer la valeur du χ^2 partiel pour chacun des axes : il vaut 8267*0,1495 = 1235,91 pour l'axe 1 (*cf.* tableau 3.6), soit 78,3 % du χ^2 total (*cf.* Tenenhaus, p. 158).

© Dunod – Toute reproduction non autorisée est un délit.

Tableau 3.6 – Tableau des valeurs propres*

Tableau des valeurs propres			
Trace de la matrice : 0,19084			
Numéro des axes	Valeur propre (lambda : λ)	Pourcentage par axe	Pourcentage cumulé
1	0,1495	78,31	78,31
2	0,0280	14,67	92,98
3	0,0088	4,65	97,63
4	0,0038	2,02	99,65
5	0,0007	0,35	100,00
total	0,1908	100,00	

* Ce traitement a été réalisé avec le logiciel SPAD.

La *seconde colonne* donne l'inertie de chaque valeur propre (symbolisée par la lettre grecque lambda : λ), laquelle indique la part de variance – l'inertie – qu'un axe factoriel donné restitue (la somme des valeurs propres est donc égale, à l'inertie totale, ou « trace »). Le ratio (exprimé en %) «λ_k / inertie totale » mesure le pourcentage de variance qu'un axe prend en charge, c'est-à-dire la quantité d'information initiale dont il rend compte à lui seul[1].

Exemple : ainsi l'axe 1 rend compte de plus des ¾ de la variance totale contenue dans le tableau croisé 6 x 11 : 0,1495/0,19084 = 78,31 %.

Les *troisième et quatrième colonnes* fournissent le pourcentage de la variance expliquée par chaque axe (respectivement en brut et en cumulé). C'est cette information qui aide à situer la valeur « explicative » de chaque axe et à évaluer la qualité du modèle retenu. Les valeurs propres sont présentées dans l'ordre décroissant de leur contribution à la variance expliquée.

L'utilisation de ces informations

Pour résumer l'information. L'analyse factorielle des correspondances est une technique statistique de réduction de l'information (*cf.* chapitre 4) : un petit nombre d'axes doit rendre compte de la plus grande part de l'inertie. C'est possible quand le lien entre les deux variables étudiées est significatif (ce qui est avéré dans l'exemple par un test d'indépendance : $\chi^2 = 1\,577{,}09$ et pour 50 degrés de liberté, la probabilité critique est inférieure à 1/10 000).

Pour construire un modèle opérationnel et valide. Les axes sélectionnés doivent rendre compte de plus de la moitié (> 50 %) de l'inertie, sinon la variance expliquée

1. Cette valeur exprime le pourcentage de relation entre les deux variables et demande à être interprété avec prudence : si la liaison est faible, l'axe rend bien compte de la forte proportion d'une liaison qui est…faible! D'où l'intérêt de vérifier préalablement l'existence d'une liaison par un test d'indépendance (*cf.* Evrard, Y et *alii*, p.431). Cependant, même si le χ^2 n'est pas significatif, l'analyse des correspondances peut révéler une structure dans les données qu'il peut être intéressant d'analyser (*cf.* Tenenhaus, p. 152).

par le modèle serait inférieure à la variance résiduelle due aux aléas : le modèle ne serait donc pas opératoire. Le choix des axes se fait en fonction de l'importance décroissante de leur contribution à la prise en charge de la variance totale.

> Dans l'**exemple des plats cuisinés**, le premier axe prend en charge plus des trois quarts de l'inertie (78,33 %) ; le suivant, seulement 1/7ième (14,67 %) soit moins de 1/5ième du poids du premier. Pratiquement, un seul axe rend compte de l'essentiel de la variance, les suivants n'ayant qu'un apport secondaire. Mais pour construire une carte perceptuelle, il faut deux axes et les deux premiers rendent compte de 93 % de la variance contenue dans le tableau croisé initial. Ce modèle, plutôt économique (un seul axe suffirait) témoigne donc d'une très forte structuration des données soumises à l'analyse.

Les procédures pour sélectionner les axes significatifs

Il existe plusieurs méthodes, statistiques ou empiriques, qui toutes conduisent à sélectionner un même nombre d'axes :

- certains logiciels (tel SPAD) proposent un test pour déterminer le nombre d'axes significatifs en s'appuyant sur la probabilité critique associée à la valeur du λ^2 calculé pour chaque axe. Dans le cas traité, un seul axe est hautement significatif ;
- la plupart des logiciels éditent un histogramme des valeurs propres et/ou une information sur les écarts de variance en passant d'un axe au suivant (*cf.* figure 3.1) ; comme la variance restituée est décroissante, un changement de concavité dans l'histogramme (ou un saut important dans les écarts de variance) correspond à une réduction importante de l'information restituée par le facteur suivant ; cette concavité (ou « coude ») définit un seuil de sélection : on ne retiendra que les axes situés avant ce seuil ;
- comme chaque modalité contribue à l'inertie totale, en cas d'indépendance entre les deux variables, l'inertie se distribuerait de manière aléatoire entre les axes ; la quantité de variance expliquée par un axe doit donc être supérieure au nombre d'axes théoriquement susceptible de rendre compte de toute l'inertie, soit le nombre de modalités de la variable la plus courte, moins une.

> Dans l'exemple traité, la variable la plus courte correspond aux modalités désignant les six modes de conservation et, par définition, comme il y a au maximum 5 axes possibles, la quantité de variance portée par un axe qui doit être dépassée (valeur seuil) vaut : 100 %/5 = 20 % ; le premier axe est bien le seul significatif (*cf.* figure 3.1).

Cependant, les critères empiriques ne sont qu'indicatifs, car c'est lors de l'analyse que la sélection finale sera faite, partant du principe qu'on ne retiendra finalement que des axes qui soient interprétables.

© Dunod – Toute reproduction non autorisée est un délit.

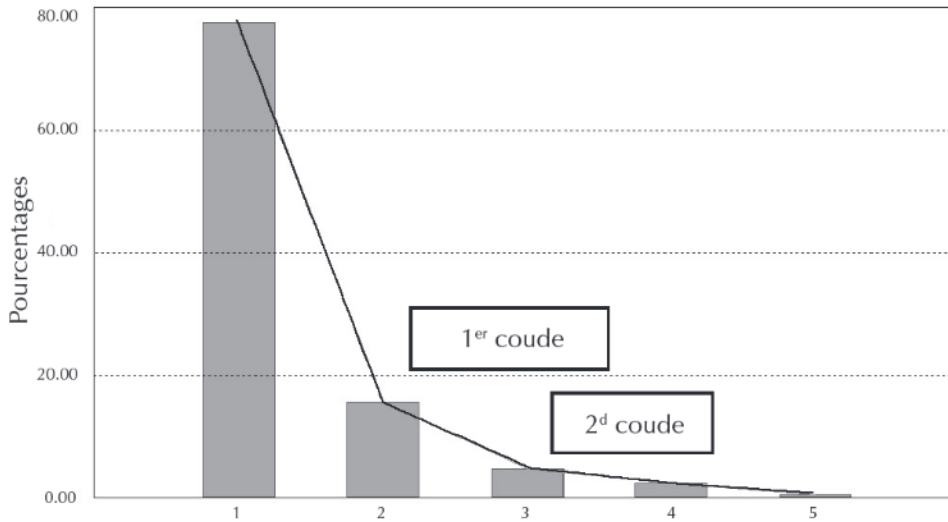


Figure 3.1 – Histogramme des cinq valeurs propres

2.2 Deuxième étape : interpréter les axes avec les informations pertinentes

Trois types d'informations permettent d'interpréter le modèle statistique construit : les contributions absolues (CTR), les coordonnées (COOR) et les contributions relatives (COS^2).

Les contributions absolues (CTR) pour identifier les axes significatifs

La valeur associée à chaque modalité dans chacun des axes correspond à la contribution de la modalité à l'inertie de l'axe : cela permet d'identifier les modalités qui ont significativement contribué à l'émergence d'un axe donné

Pour chaque axe, la somme des contributions des modalités vaut 1 puisqu'elle correspond à la variance d'un axe normé par construction (centré et réduit, de moyenne 0 et d'écart-type 1) – pour une meilleure lisibilité, la contribution peut être multipliée par 100 pour s'apparenter à un pourcentage (*cf.* tableaux 3.7 et 3.8).

Tableau 3.7 – Tableau des contributions absolues pour les « attributs » (x 100)*

Modalités des variables :	Poids relatif (% des modalités)	Distance à l'origine	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5
			CTR ou Contributions absolues x 100				
ATTRIBUTS							
plaisir	9,84	0,10560	5,19	6,15	6,95	5,59	**12,85**
qualité/prix	8,06	0,10918	0,14	**23,32**	6,55	**38,21**	1,34
originalité	9,34	0,07887	4,01	2,50	0,89	**14,69**	4,06
bonne qualité	6,49	0,17992	3,23	3,50	**66,36**	0,34	0,21
saveur	8,89	0,14783	8,05	3,02	2,78	0,01	3,98
stockage	8,00	0,65461	**28,75**	**30,24**	1,20	**20,27**	**9,22**
praticité	11,06	0,27246	**18,15**	8,84	3,75	0,02	**28,68**
rapidité	11,83	0,30321	**19,93**	**19,43**	0,08	**14,34**	**10,89**
naturel	8,15	0,15307	7,68	0,77	6,73	3,35	8,32
recettes	8,69	0,05491	2,74	1,37	1,33	2,54	**11,49**
diversité	9,64	0,03940	2,13	0,84	3,37	0,63	8,97
total	100,00		100,00	100,00	100,00	100,00	100,00
PRODUITS							
maison	20,90	0,32540	**41,51**	**18,67**	6,24	4,37	8,30
conserve	16,97	0,18932	**18,76**	**6,48**	**22,52**	1,88	**33,40**
traiteur	17,45	0,14464	5,39	**57,26**	8,30	11,32	0,28
surgelé	22,95	0,02848	0,02	6,52	**41,13**	**26,05**	3,33
déshydraté	9,88	0,51720	**33,88**	0,03	1,14	0,58	**54,47**
frais sous vide	11,84	0,06527	0,44	11,04	**20,66**	**55,79**	0,23
total	100,00		100,00	100,00	100,00	100,00	100,00

* Calculs effectués avec le logiciel SPAD.

La question se pose cependant de savoir à partir de quelle valeur la contribution d'une modalité peut être considérée comme déterminante de l'émergence d'un axe. Dans la pratique, cette valeur doit être supérieure au ratio 100/q (où q est le nombre de modalités de la variable et 100 la variance totale de l'axe) ; en effet, cette valeur correspond à ce que vaudrait l'inertie dans chaque modalité si les deux variables étaient indépendantes.

Dans l'exemple traité, ce ratio vaut 9,09 pour les attributs (100/11) et 16,67 pour les produits (100/6). Les modalités supérieures à ces valeurs sont en italiques. Soit, en ordre décroissant des contributions pour chaque axe :

- les modalités {*stockage, rapidité, praticité*} et {*maison, déshydraté* puis *conserve*} ont généré l'axe 1 ;
- les modalités {*stockage, qualité/prix, rapidité*} et {*traiteur,* puis *maison*}, l'axe 2 ;
- les modalités {*bonne qualité*} et {*surgelé, conserve*}, l'axe 3 ;
- les modalités {*qualité/prix, stockage, originalité, rapidité*} et {*frais sous vide, surgelé*}, l'axe 4 ;

© Dunod – Toute reproduction non autorisée est un délit.

– quant au cinquième axe, il est généré par cinq attributs différents associés à deux produits : *déshydraté* et *conserve*.

Mais il ne faut pas oublier que le premier axe est, ici, le seul significatif : il rend compte de plus des ¾ de la variance totale, les suivants n'ont donc qu'un intérêt anecdotique.

Les coordonnées (COOR) des modalités pour interpréter les axes

Si les contributions absolues (CTR) ont permis d'identifier les modalités significatives de l'existence d'un axe, elles ne contribuent que partiellement à en définir le sens. Il faut compléter cette information avec celle contenue dans la matrice des coordonnées des modalités. En effet, ce sont ces coordonnées qui vont permettre de donner toute sa signification à un axe. Comme elles sont affectées d'un signe, positif (+) ou négatif (-), il est donc possible de situer sur un axe, parmi les modalités significatives, celles qui sont voisines et celles qui sont opposées, comme l'illustre le tableau 3.8 ; le sens d'un axe naît de cette mise en perspective.

Illustration pour la variable « Attributs » : sur l'axe 1, les trois modalités significatives sont voisines (praticité et rapidité se superposent même) ; sur l'axe 2, qualité/prix et stockage sont voisins et s'opposent à rapidité ; sur l'axe 4, saveur et praticité se superposent et sont voisines de qualité/prix, mais s'opposent à originalité.

Tableau 3.8 – Coordonnées (COOR) des « attributs » et des « produits »

Modalités des variables :	Poids relatif	Distance à l'origine	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5
			coordonnées des modalités (COOR)				
ATTRIBUTS							
plaisir	9,84	0,10560	0,28	0,13	0,08	-0,05	-0,03
qualité/prix	8,06	0,10918	0,05	-0,28	0,09	0,14	0,01
originalité	9,34	0,07887	0,25	0,09	0,03	-0,08	0,02
bonne qualité	6,49	0,17992	0,27	-0,12	-0,30	-0,01	-0,01
saveur	8,89	0,14783	0,37	0,10	0,05	0,00	-0,02
stockage	8,00	0,65461	-0,73	-0,33	0,04	-0,10	-0,03
praticité	11,06	0,27246	-0,49	0,15	-0,05	0,00	0,04
rapidité	11,83	0,30321	-0,50	0,22	-0,01	0,07	-0,02
naturel	8,15	0,15307	0,38	-0,05	-0,09	0,04	-0,03
recettes	8,69	0,05491	0,22	-0,07	0,04	-0,03	0,03
diversité	9,64	0,03940	0,18	-0,05	0,05	0,01	0,02
PRODUITS							
maison	20,90	0,32540	0,54	-0,16	0,05	0,03	-0,02
conserve	16,97	0,18932	-0,41	-0,10	0,11	0,02	0,04
traiteur	17,45	0,14464	0,21	0,30	0,06	-0,05	0,00
surgelé	22,95	0,02848	-0,01	-0,09	-0,13	-0,07	0,01
déshydraté	9,88	0,51720	-0,72	-0,01	0,03	-0,02	-0,06
frais sous vide	11,84	0,06527	-0,07	0,16	-0,12	0,13	0,00

Le tableau 3.9 synthétise l'information nécessaire pour l'interprétation des axes, sans négliger leur poids (inertie) ni la contribution des modalités. Cependant, pour finaliser le modèle, il faut encore apprécier la capacité des axes à rendre compte de la variance des modalités des variables.

Tableau 3.9 – Synthèse des modalités significatives pour l'interprétation des axes

Variables :		« attributs »				« objets »			
Axe	Inertie %	Valence +	CTR	Valence -	CTR	Valence +	CTR	Valence -	CTR
1	78,31	stockage rapidité praticité	28,8 19,9 18,2			déshydraté appertisé	33,9 18,8	maison	41,5
2	14,67	rapidité	19,4	stockage qualité/prix	30,2 23,3	traiteur	57,3	maison	18,7
3	4,65	qualité	66,4			surgelé frais/vide	41,1 20,7	appertisé	22,5
4	2,02	stockage originalité	20,3 14,7	qualité/prix rapidité	38,2 14,3	surgelé	26,1	frais/vide	55,8
5	0,35	praticité recettes	28,7 11,5	plaisir rapidité	12,8 10,9	conserve	33,4	déshydraté	54,5

L'apport des contributions relatives (COS^2) pour évaluer le modèle

Le cosinus-carré, encore appelé *contribution relative*, mesure la part de variance d'une modalité qui est prise en charge – ou « expliquée » – par un axe. Pour une modalité donnée, la valeur des cosinus-carré cumulée sur l'ensemble des axes vaut 1 (tableau 3.10). Cette information permet donc de savoir si, pour rendre compte de la plus grande partie de la variance d'une modalité, un seul axe suffit (tel l'axe 1 pour les modalités *naturel* et *plat déshydraté*) ou non (tels les axes 1 et 3 pour *bonne qualité* et les axes 2, 3 et 4 pour *plats sous vide*).

Ainsi, comme le modèle final ne contient que les axes significatifs (et non tous), il faut alors vérifier que l'ensemble des seuls axes retenus prend bien en charge :
- la plus grande partie de la variance de chaque modalité (cette proportion devrait être la plus grande possible et, au minimum, supérieure à 50 %) ;
- et cela pour le plus grand nombre de modalités, sinon toutes idéalement.

Exemple des « plats cuisinés ». L'axe 1 prend bien en charge la variance de 9 des 11 attributs, mais seulement celle de 3 des 6 produits : il est logique qu'un facteur dont la valeur propre est élevée rende compte de la variance d'un grand nombre de modalités puisque ce sont ces modalités qui l'ont généré. Ce premier facteur explique à lui seul entre 75 % et 91 % de leur variance (un seuil pragmatique de 60 % correspond à une prise en charge satisfaisante). Un modèle en deux facteurs rend compte de la quasi-totalité de la variance des items (à l'exception de l'item *bonne qualité* : 0,49 et des deux modes de *conservation frais sous vide* : 0,48 et surtout *surgelé* : 0,29 qui nécessiteraient de retenir aussi le troisième facteur). La plus grande partie de l'inertie totale devant être prise en charge par le plus petit nombre d'axes possibles (principe d'économie), un modèle à deux axes peut être utile ici pour rendre compte au mieux de la variance du maximum de modalités : un tel modèle explique 93 % de l'inertie totale.

© Dunod – Toute reproduction non autorisée est un délit.

Tableau 3.10 – Tableau des contributions relatives ou cosinus-carré

Modalités des variables	Poids relatif	Distance à l'origine	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	cosinus² tous axes : Somme	cosinus² axes 1+2 : Somme
ATTRIBUTS									
plaisir	9,84	0,10560	0,75	0,17	0,06	0,02	0,01	1,00	0,91
qualité/prix	8,06	0,10918	0,02	0,74	0,07	0,17	0,00	1,00	0,76
originalité	9,34	0,07887	0,81	0,09	0,01	0,08	0,00	1,00	0,91
bonne qualité	6,49	0,17992	0,41	0,08	0,50	0,00	0,00	1,00	0,49
saveur	8,89	0,14783	0,91	0,06	0,02	0,00	0,00	1,00	0,98
stockage	8,00	0,65461	0,82	0,16	0,00	0,01	0,00	1,00	0,98
praticité	11,06	0,27246	0,90	0,08	0,01	0,00	0,01	1,00	0,98
rapidité	11,83	0,30321	0,83	0,15	0,00	0,02	0,00	1,00	0,98
naturel	8,15	0,15307	0,92	0,02	0,05	0,01	0,00	1,00	0,94
recettes	8,69	0,05491	0,86	0,08	0,02	0,02	0,02	1,00	0,94
diversité	9,64	0,03940	0,84	0,06	0,08	0,01	0,02	1,00	0,90
PRODUITS									
maison	20,90	0,3254	0,91	0,08	0,01	0,00	0,00	1,00	0,99
appertisé	16,97	0,1893	0,87	0,06	0,06	0,00	0,01	1,00	0,93
traiteur	17,45	0,1446	0,32	0,63	0,03	0,02	0,00	1,00	0,95
surgelé	22,95	0,0285	0,01	0,28	0,56	0,15	0,00	1,00	0,29
déshydraté	09,88	0,5172	0,99	0,00	0,00	0,00	0,01	1,00	0,99
frais/vide	11,84	0,0653	0,08	0,40	0,24	0,28	0,00	1,00	0,48

L'exploitation de toutes ces informations – coordonnées, contributions absolues et relatives – telle que présentée dans le tableau 3.9 prépare l'interprétation des axes. Mais elle prépare aussi l'édition des cartes perceptuelles : les cosinus-carré permettent en effet de contrôler la qualité de la représentation des modalités dans les plans édités et limitent ainsi les risques d'interprétation erronée ou abusive.

2.3 Troisième étape : analyser et interpréter les résultats du modèle

La validité opérationnelle du résultat ne vaut que par la qualité et la pertinence des informations qui ont été introduites dans le traitement, sachant que le modèle factoriel ne rend compte que des informations qui ont été intégrées dans l'analyse. Aussi, il importe de distinguer entre ce qui relève de la structuration des données exploitées (algorithme) et ce qui relève de l'interprétation que peut en donner l'analyste. D'où la pertinence de distinguer entre :

– les axes significatifs du point de vue statistique prenant en charge l'essentiel de l'information (l'inertie) contenue dans le tableau croisé ; ces axes constituent la base du modèle retenu (les deux premiers facteurs dans l'exemple traité) ;

– les axes « secondaires » qui ne rendent compte que d'une toute petite partie de l'inertie, mais dont l'émergence peut nous révéler une certaine structuration des données ou nous interroger sur la pertinence de l'étude conduite et ses conséquences managériales. Ces facteurs ne rentrent pas dans le modèle retenu (soit les facteurs 3 à 5 dans l'exemple traité).

L'interprétation des axes significatifs

L'objectif est de rechercher ce qu'il y a de commun ou de transversal à toutes les modalités significativement liées à un axe (et identifiées par les contributions absolues) afin de dégager sa signification propre et lui trouver l'appellation adaptée résumant au mieux le contenu qu'il porte (en évitant tout nominalisme).

Application à la perception des plats cuisinés.

– L'axe 1 (78 % de l'inertie totale) est lié au mode de transformation du produit : il oppose l'offre industrielle à l'offre domestique. Il associe significativement (tableau 3.9) les conserves « déshydratée » et « appertisée » avec une certaine commodité d'usage (durée du *stockage*, *rapidité* et *facilité d'emploi*) qu'il oppose à la conserve « maison » : il renvoie à la nature du plat cuisiné qui est le résultat soit d'une transformation industrielle, soit d'une transformation artisanale. Par ailleurs, cet axe est focalisé sur les seuls avantages du produit industriel (les qualités du produit maison – modalités *naturelles*, *saveur* – ne sont lui pas significativement liées).

– L'axe 2 (environ 15 % de la variance totale et donc très secondaire comparé au premier axe) complète l'information précédente. En effet, son introduction améliore le taux de prise en charge de la variance totale (93 %) et il intègre le rôle de modalités mal prises en compte par le premier facteur. C'est cette complémentarité qui incite à le retenir, car il précise les caractéristiques de la dimension artisanale : il oppose « plat maison » reconnu pour son avantage *qualité/prix* et sa facilité de *stockage* au « plat traiteur » dont l'avantage est d'être d'usage *rapide*.

L'apport des axes secondaires

Un axe associé à un nombre réduit de modalités (voire une seule) présente donc une certaine spécificité qui l'exclut du modèle général. Le caractère « spécifique » d'un axe conduit à s'interroger sur la pertinence des modalités dont il rend compte :

– soit leur pertinence est douteuse : elles relèvent d'un autre champ ou supportent plusieurs significations (ambivalence sémantique) ;
– soit leur pertinence est avérée, mais leur contenu est insuffisamment approfondi (item isolé dont l'idée aurait mérité d'être explicitée ou développée avec d'autres modalités ; offre particulière dont la nature devrait être précisée par l'introduction de produits voisins ; etc.). Il faut donc rechercher les raisons à cette « spécificité ».

Application à la perception des plats cuisinés.

– L'axe 3 – 4,6 % de la variance expliquée – oppose le produit conservé frais (surgelé et sous vide) à la boîte de conserve appertisée sur la base d'un différentiel de qualité au bénéfice du frais. En fait, cela complète la compréhension du premier axe.

© Dunod – Toute reproduction non autorisée est un délit.

– L'axe 4 – 2 % de la variance expliquée – concerne le produit frais industriel et marque une différence entre le surgelé (qui bénéficie d'une conservation plus longue avec une offre plus originale) et le frais sous vide (reconnu pour son rapport qualité/prix et la rapidité d'emploi).
– L'axe 5 (insignifiant en regard de la variance expliquée : 0,35 %) a été généré par cinq modalités attributs, dont deux lui sont spécifiques : plaisir et recettes. Comparé au contenu des quatre axes précédents qui éclairent sur la perception des différents produits, celui-ci n'apporte rien de plus et relève d'une autre logique : ces deux attributs traduisent en fait une forte spécificité. En effet, alors que la majorité des items renvoie aux caractéristiques, aux qualités intrinsèques des produits :
+ l'item *plaisir* réfère plutôt au consommateur et à une certaine connotation psychologique : son contenu n'est pas de même nature que les autres items et cette modalité n'aurait sans doute pas dû être introduite dans l'étude ;
+ l'item *recettes* est ambigu : il peut être compris comme « diversité de l'offre », ce qui le rapprocherait de *11- diversité des goûts* ; mais aussi comme recette au sens technique du terme, faisant du produit une préparation culinaire travaillée avant d'être servie ; il peut être aussi compris au premier degré : un conditionnement qui contient des idées de recettes.

L'exemple de l'axe 5 permet de soulever quelques problèmes méthodologiques liés au choix et à la formulation de certaines modalités référant aux attributs : on peut se demander s'il était opportun de les introduire dans l'étude et, à considérer ces résultats, s'ils ne pourraient pas être supprimés pour effectuer une nouvelle analyse avec seulement les neuf attributs restant.

Les axes 3 et 4, en permettant d'affiner la perception qu'en ont les répondants, éclairent sur le positionnement des produits selon leur mode de conservation, bien que ce complément d'information reste assez marginal par rapport au trait de différenciation majeur (plutôt trivial ici puisqu'il oppose le produit industriel au produit artisanal).

Si l'objectif de cette analyse est de créer des variables composites résumant l'information pour effectuer d'autres types de traitements (une typologie par exemple), un seul axe s'impose : le premier. Mais si l'objectif est de comprendre la perception des consommateurs, l'étude des axes secondaires peut s'avérer utile.

La figure 3.2, qui structure les associations dans les différents facteurs, montre la cohérence de l'ensemble du système de référence utilisé pour comparer les divers plats cuisinés : l'ensemble est clairement organisé autour d'un axe qui dissocie les plats industrialisés perçus comme ayant subi une transformation et les plats peu transformés perçus comme plus naturels. C'est, *a posteriori*, une vérification de la qualité et de la pertinence du modèle conceptuel utilisé (que les modalités des variables traduisent). Mais cette construction n'est pas toujours évidente ; c'est précisément l'un des avantages de l'analyse des correspondances que de pouvoir élaborer des cartes perceptuelles pour aider l'interprétation.

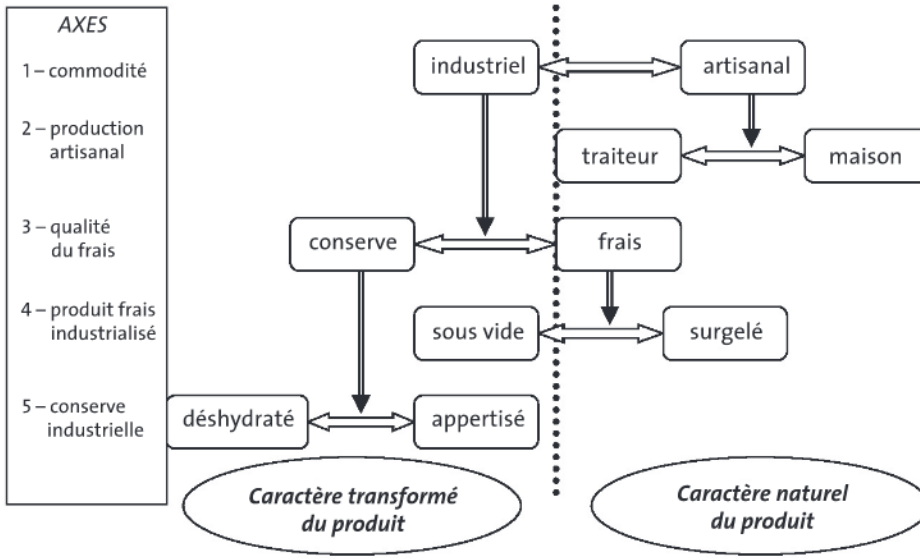


Figure 3.2 – Structuration de l'univers d'évaluation des plats cuisinés

2.4 Quatrième étape : construire les cartes perceptuelles

L'analyse factorielle des correspondances permet de visualiser simultanément sur une carte l'espace des lignes (les objets) et celui des colonnes (les attributs) pour donner une visibilité immédiate des relations de voisinage ou d'opposition entre les modalités.

Chaque plan croise deux axes dont le point d'intersection a valeur d'origine. Cette origine correspond au centre de gravité du nuage de points généré par toutes les observations obtenues sur les deux variables étudiées ; il représente en quelque sorte le profil moyen des variables. En conséquence, une modalité proche de ce point d'origine (l'intersection des axes) ne se distingue pas de la masse et n'est pas significative de l'axe ; par contre plus une modalité s'éloigne de l'origine et plus elle devient significative de l'axe sur lequel elle est projetée.

L'apport des contributions relatives (COS²) dans la lecture des axes

La lecture directe des cartes perceptuelles n'est pas sans risque, car la position des points dans un plan peut ne pas être représentative de leur position dans l'espace multidimensionnel initial. La matrice des cosinus-carré – ou *contributions relatives* – contient l'information utile pour apprécier la qualité de représentation des points-modalités dans chaque plan croisant deux axes factoriels.

© Dunod – Toute reproduction non autorisée est un délit.

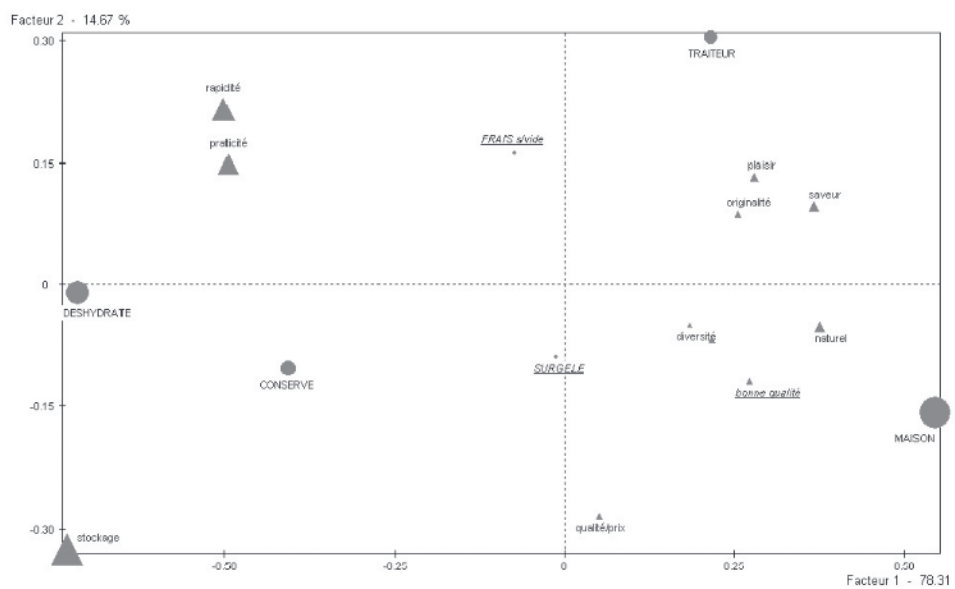
Géométriquement, le cosinus-carré mesure la qualité de la projection du point original sur le plan : il correspond à la distance entre cette projection et l'origine de l'axe. Variant de 0 à 1, la qualité de la projection est maximale pour 1 (superposition exacte entre le point original et sa projection). Cette information est indispensable pour apprécier la proximité réelle entre deux modalités : les deux points qui les représentent sur un axe peuvent être très proches dans le plan alors qu'ils sont en réalité très éloignés, voire opposés dans l'espace.

Ce défaut de perspective peut donc être la source d'une interprétation erronée, comme l'illustre la figure 3.3 qui présente la projection de l'ensemble des modalités contenues dans le tableau croisé initial (alors qu'elles ne sont pas toutes significatives, ni bien représentées).

Application à la perception des plats cuisinés.

– Une lecture directe des résultats sur le plan peut conduire à superposer *maison* et *traiteur* sur l'axe 1 (horizontal) avec l'ensemble des sept items inclus dans la zone qui les sépare ; si l'association est cohérente ici, elle ne correspond cependant pas exactement aux seules données qui sont significatives.

– Procéder de même sur l'axe 2 (vertical) conduit à superposer *traiteur* et *frais sous vide* avec *rapidité* et *originalité*, en opposition avec *maison* et *conserve* qui, en plus, sont associés à *bonne qualité*. Même si en apparence la logique est préservée, les écarts de position (cosinus² insuffisants) induisent des associations qui n'ont pas lieu d'être : *plaisir* et *recettes* ne sont pas significatifs sur les deux premiers axes et ne contribuent qu'à l'émergence de l'axe 5, alors qu'ils sont bien placés sur le plan issu des deux premiers axes (*cf.* dernière colonne du tableau 3.10).



Source : Obtenue avec SPAD

Figure 3.3 – Carte perceptuelle sur les deux premiers axes

Remarque

La taille des idéogrammes (rond, triangle) croît avec la valeur des contributions ; les trois modalités mal représentées sur ce plan sont en italique et soulignées.

D'autres causes d'erreur d'interprétation liée à la lecture directe d'un plan

– Compte tenu de la métrique du graphique (qu'il faut considérer avec attention), les pôles opposés des axes ne sont pas nécessairement symétriques par rapport au point d'origine (croisement des axes), compte tenu de leur longueur (c'est le cas de l'axe 1 dans la figure 3.3). Il faut donc vérifier si la position des points les plus extrêmes se détache significativement de la position moyenne définie par l'origine des axes (le centre de gravité du nuage de points).
– Le profil de distribution d'une modalité particulièrement originale ou marginale peut générer un axe avec une valeur propre élevée. Une telle situation est repérable car elle peut entraîner un déséquilibre dans la longueur des axes sur le graphique. Selon l'importance de cette modalité en regard des objectifs de l'étude réalisée, il faut se demander s'il convient de la supprimer avant d'effectuer une nouvelle analyse. La même question doit être posée quand une modalité n'est pas prise en charge par les principaux axes du modèle et lorsqu'aucun d'eux ne rend significativement compte d'une importante fraction de sa variance ; c'est précisément ce que l'on peut observer avec la modalité *frais sous vide* dans l'exemple traité (*cf.* tableau 3.10).

Les différents types de projection et leur mode d'utilisation

Dans une carte perceptuelle, on peut traduire la proximité entre les modalités des variables de deux manières, soit en utilisant une représentation standard – la plus usitée – dans laquelle les modalités en ligne et en colonne jouent un rôle identique : elles sont présentées sous forme de points. Une autre représentation, appelée barycentrique (ou « biplot » dans les logiciels d'origine anglo-saxonne), projette les modalités d'une variable dans l'espace originel de l'autre variable ; les modalités de la variable source sont présentées sous forme de point quand celles de la variable « importée » le sont sous forme de vecteur. Ces deux modes de présentation répondent à des préoccupations complémentaires :

– *L'étude des proximités entre modalités appartenant à une même variable :* la représentation standard donne une image fidèle des proximités. Quel que soit le quadrant considéré dans le plan, quand deux modalités d'une même variable sont bien représentées (Cosinus-carré élevé), on peut donc conclure :
  – qu'elles se ressemblent si elles sont voisines ;
  – qu'elles ont peu ou pas de points communs si elles sont opposées.

© Dunod – Toute reproduction non autorisée est un délit.

Application à la perception des plats cuisinés.
Ainsi, *rapidité* et *praticité*, très bien représentées sur le plan (cosinus² cumulé de 0,98 sur les deux axes du plan) et dans le même quadrant, sont assurément voisins (schéma 3.3) ; *rapidité* et *saveur* sont opposés : le contenu de ces deux attributs est très différent. Un raisonnement identique permet de relever le voisinage entre le déshydraté et la conserve appertisée ; ou l'opposition entre le déshydraté et le plat maison qui n'ont effectivement rien de commun.

– *L'étude des proximités entre modalités appartenant à des variables différentes* : si les modalités sont bien représentées (cosinus-carré élevé), la représentation standard reste adaptée, mais dans les conditions suivantes :
  – si une modalité colonne (« objet ») est dans le même quadrant qu'une modalité ligne (« attribut »), l'objet possède l'attribut (et inversement) ;

Les produits *déshydratés* et *appertisés* partagent la qualité *stockage*.

  – si une modalité colonne (« objet ») est dans le quadrant diagonalement opposé à celui d'une modalité ligne (« attribut »), cet objet ne possède pas cet attribut ;

Le produit *maison* ne possède pas les caractéristiques *rapide* et *pratique*.

  – si une modalité colonne (« objet ») et une modalité ligne (« attribut ») sont dans des quadrants adjacents, il n'est pas possible de conclure : la représentation standard n'est pas adaptée à cette situation.

On ne peut assurer que la conserve *appertisée* est *pratique* ou que la conserve *maison* n'est *pas stockable* sans devoir le vérifier d'une autre manière.

– *Lorsque les modalités sont dans des quadrants adjacents* dans la représentation standard, il fait recourir à la représentation barycentrique (ou « biplot ») pour analyser en toute sécurité les liens de proximité ou d'opposition qui existent entre les modalités des deux variables étudiées. Si ces modalités sont bien représentées (cosinus-carré élevé), la méthode consiste à considérer l'angle que fait la projection des deux modalités en passant par l'origine des axes :
  – si l'angle formé est aigu (inférieur à 90°) : il y a une relation de voisinage entre les modalités des deux variables (figure 3.4 : trait plein) ;

C'est le cas de *praticité* et de *conserve appertisée* : ces deux modalités étant bien représentées (cos² cumulé supérieur à 0.90), on peut donc bien dire que la conserve est perçue comme « pratique » (et aussi « rapide »).

  – si l'angle formé est obtus (supérieur à 90°) : il y a une relation d'opposition entre les modalités des deux variables (figure 3.4 : trait pointillé).

C'est le cas de *stockage* et de *conserve maison* : on peut donc dire que ce produit est perçu comme ne se conservant pas longtemps, ou pas fait pour être stocké.

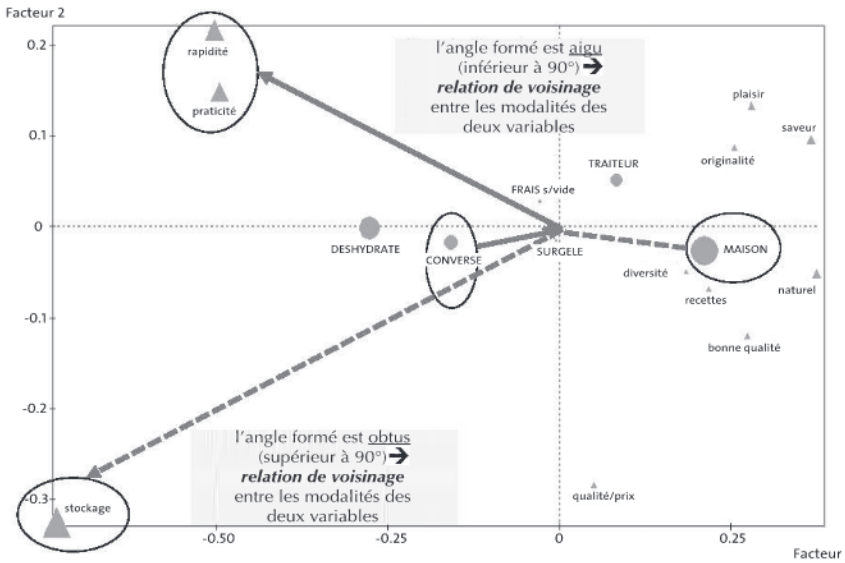


Figure 3.4 – Représentation barycentrique ou « biplot » : les relations de voisinage et d'opposition

Les risques d'erreur liés à l'interprétation des axes

Déjà évoqués, on peut les résumer pour clore cette partie :

– *La lecture directe* des résultats sur une carte perceptuelle peut mener à vouloir associer des modalités en apparence voisines alors qu'en réalité elles ne le sont pas (exemple : « saveur » et « naturel » dans le cas traité). Il faut donc commencer par identifier les seules modalités significatives pour chaque paire d'axe (les contributions absolues), puis apprécier la qualité de leur projection dans le plan (les cosinus-carré ou contributions relatives) et vérifier la qualité de l'interprétation avec une projection barycentrique.

– *Le psychologisme*. Donner une dimension ou une interprétation psychologique à un axe, qui est de nature fondamentalement statistique, n'est pas pertinent. En effet, la quantité de variance « expliquée » par les axes (mesurée par les valeurs propres) ne préjuge en rien de l'importance que le contenu de cet axe peut avoir pour les répondants. Il serait par exemple erroné d'affirmer que l'origine industrielle ou artisanale d'un plat cuisiné (premier axe) détermine le choix des consommateurs : l'AFC ne permet pas une telle conclusion. Par contre, ce qu'elle montre, c'est que l'information introduite dans l'étude est focalisée (à considérer le nombre et la nature des modalités) sur une thématique spécifique, qui concerne bien les caractères industriel et artisanal des conserves.

– *Le nominalisme.* Il est usuel de donner un nom à chaque axe pour résumer au mieux l'information commune et transversale à l'ensemble des modalités qui le constituent. Mais le risque est de retenir une appellation en oubliant la réalité du

© Dunod – Toute reproduction non autorisée est un délit.

contenu, en se forgeant ainsi des représentations qui peuvent s'éloigner de l'information originelle. Or, ce sont ces représentations qui risquent de guider la réflexion, voire les décisions managériales. Il faut donc se méfier du pouvoir des mots dont le choix relève parfois plus de la volonté de séduire que du souci de transcrire objectivement l'information contenue dans l'axe.

Aussi, pour toutes ces raisons, il importe de ne pas se limiter aux seules cartes fournies et à la dénomination des axes. Pour la réflexion, il faut toujours se référer aux tableaux contenant les résultats statistiques : matrices des valeurs propres, des contributions absolues, des coordonnées et des contributions relatives.

L'ESSENTIEL

■ ***Finalité de l'analyse factorielle des correspondances simples (AFC)***

L'AFC permet d'étudier un tableau de contingence dont les deux variables non-métriques qu'il associe comportent un nombre important de modalités.

■ ***Une démarche d'analyse à conduire avec rigueur***

Pour garantir l'opérationnalité de l'usage de l'outil statistique et l'objectivité des conclusions auxquelles il conduit, l'analyse des résultats doit se dérouler en plusieurs phases :

– identifier les axes significatifs à partir du tableau des valeurs propres ;

– définir le contenu des axes avec les seules modalités dont la contribution absolue est significative, et en intégrant leur position de voisinage ou d'opposition à partir de leurs coordonnées ;

– vérifier la capacité du modèle retenu (choix du nombre d'axes significatifs) à prendre en charge la variance du maximum de modalités et évaluer sa validité globale (fraction de l'inertie totale du tableau de contingence prise en charge) en analysant conjointement les contributions absolues et relatives ;

– opter éventuellement pour des représentations graphiques en complément de l'analyse indispensable des données chiffrées.

■ ***Faire un bon usage des graphiques***

Comme la lecture directe des plans-graphiques projetant les modalités sur des couples d'axes peut entrainer des erreurs d'interprétation, il faut respecter très scrupuleusement les règles de lecture, particulièrement pour analyser les liens entre modalités qui appartiennent à des quadrants adjacents dans le plan de projection des axes.

Chapitre

4

Comment simplifier l'information avec l'analyse factorielle

OBJECTIFS

- Connaître les différentes méthodes d'analyse factorielle métrique (ACP) et non métrique (ACM) et leurs applications pour choisir le modèle adapté à la nature des données et au type de problème à résoudre.
- Proposer une méthodologie permettant, par étapes successives, d'obtenir un modèle factoriel pertinent et opérationnel.
- Fixer quelques recommandations pour éviter les interprétations abusives et permette une bonne utilisation, tant technique que managériale, des méthodes d'analyse factorielle.

SOMMAIRE

SECTION 1 L'analyse factorielle des données métriques
SECTION 2 L'analyse factorielle des données non-métriques

Les méthodes d'analyse factorielle étudient les relations qui existent entre des ensembles de variables de même statut, c'est-à-dire entre lesquelles il n'y a pas de hiérarchie établie comme dans les méthodes prédictives et causales. Leur but est de regrouper dans un même ensemble structuré – indifféremment appelé « axe », « facteur », « composante » ou « dimension » – toutes les variables liées entre elles : il y a donc autant de facteurs que d'ensembles de variables significativement liées. Mais si toutes les variables sont réellement indépendantes (sans relation), il y a autant de facteurs significatifs que de variables initiales.

La nature des variables conduit à distinguer deux types de modèles : l'analyse factorielle sur données non-métriques (analyse des correspondances, simples ou multiples) et l'analyse factorielle sur données métriques (analyse en composantes principales ou analyse en facteurs communs et spécifiques). La technique d'extraction des facteurs est la même pour tous les modèles, à savoir l'analyse en composantes principales. La différence concerne la manière de calculer les liens entre variables : la corrélation ou la covariance quand les variables sont métriques, une distance fondée sur le Chi-deux quand les variables ne sont pas métriques.

Une analyse factorielle est qualifiée d'*exploratoire* quand elle est utilisée pour identifier les liens possibles entre toutes les variables étudiées ; elle est qualifiée de *confirmatoire* quand son objet est la validation d'hypothèses préalablement formulées sur les liens spécifiques qui peuvent exister entre groupes de variables prédéfinis (ce modèle dépasse le cadre de cet ouvrage).

Quel que soit le modèle d'analyse factorielle *exploratoire*, la méthodologie reste la même, à savoir :

1. *Vérifier les conditions d'application de la méthode factorielle*
   - Contrôler la pertinence des variables
   - Vérifier que les données sont factorisables
2. *Construire le modèle factoriel proprement dit*
   - Choisir la méthode d'extraction des facteurs
   - Construire la matrice des facteurs
   - Décider du nombre de facteurs
3. *Valider la pertinence et interpréter les résultats du modèle retenu*
   - Effectuer ou non une rotation
   - Interpréter le contenu des facteurs

Mais, avant de présenter les méthodes d'analyse factorielle métrique (section 1) puis l'analyse factorielle des correspondances multiples (section 2), il importe de présenter les principaux objectifs de ces outils.

Au-delà de la mise en évidence des liens existant entre certaines variables (ou leurs modalités), l'analyse factorielle exploratoire, qu'elle soit ou non métrique, répond à un double objectif de *réduction* et de *structuration* des données.

– **L'objectif de réduction** consiste à condenser l'information initiale dans quelques variables composites qui remplacent les variables d'origine, tout en préservant l'essentiel de l'information qu'elles contiennent : ce sont les facteurs. Ainsi, ils permettent de disposer d'un nombre plus restreint de variables, généralement indépendantes, qui peuvent être utilisées dans d'autres applications statistiques.

> **Exemple.** Dans la construction d'un questionnaire, une pré-enquête peut aider à repérer quels sont les items qui se trouvent associés dans un même facteur : ce qui signifie qu'ils correspondent à un même contenu ou une même idée. Cet usage permet d'évaluer le degré de redondance pour décider s'il convient de supprimer certains items ou, au contraire, d'utiliser deux formulations d'une même idée pour construire deux questionnaires alternatifs.

– **L'objectif de *structuration*** conduit à répartir les variables initiales entre les différents facteurs : cela aide à comprendre quelle est l'information commune ou transversale à ces groupes de variables et cela révèle l'existence d'une structure latente ayant sa propre signification.

> On utilise souvent l'analyse factorielle pour construire et comprendre la structure d'une notion, comme le concept d'implication envers une catégorie de produit par exemple, décomposé en cinq facettes (intérêt, plaisir, signe social, importance du risque et probabilité de mauvais choix). Aussi, l'analyse factorielle est souvent utilisée dans cette perspective pour construire des échelles de mesure des attitudes (qui pourront être validées par une analyse confirmatoire).

Quand les variables qualifient des individus, l'analyse permet d'identifier le système – comportemental, intellectuel ou affectif – qui les gère. Ces résultats sont alors utilisés pour comparer, de manière économique, les individus sur des critères synthétiques.

L'information traitée peut aussi caractériser des produits, soit de manière objective (quantités vendues, prix de vente, parts de distribution, de linéaire, etc.), ou subjective (utilité, efficacité, image perçue, etc.) où les résultats sont interprétés en termes d'image ou de système de représentation contribuant à l'étude du positionnement.

Enfin, la qualité des résultats obtenus et leur opérationnalité en regard des décisions à prendre dépend de la qualité des variables initiales introduites dans l'analyse. Deux critères doivent être considérés :

– *L'exhaustivité* des notions, concepts et autres informations utiles pour traiter le problème posé. Si une donnée fondamentale a été oubliée ou omise, c'est la validité du modèle qui est en cause. L'étude ne vaudra que ce que vaut l'information introduite : une analyse factorielle ne peut pas signaler qu'elle repose sur des informations peu pertinentes ou partielles !
– *La complémentarité*. L'analyse repose sur les liens supposés exister entre les variables. Cela implique qu'une notion ou un concept soit abordé sous divers points de vue complémentaires ; sinon, réaliser une analyse factorielle ne se justifie guère.

© Dunod – Toute reproduction non autorisée est un délit.

Section 1

L'ANALYSE FACTORIELLE DES DONNÉES MÉTRIQUES

Cas d'illustration

« Sirop de fruits »

Refresh Ltd envisage de lancer sa marque de sirop de fruits pour adultes sur le marché français où plusieurs marques nationales s'affrontent déjà. À cette fin, elle réalise un test d'accueil comparatif : sur un ensemble de foyers consommateurs réguliers de sirop de fruits des marques françaises – marques nationales et marques de la distribution – recrutés par l'intermédiaire d'un panel « *single source* », elle dépose l'échantillon de sa marque pour essai, accompagné d'une diffusion de publicités ciblées qui explicite le positionnement de cette nouvelle offre : sirop de qualité « haut de gamme » plutôt destiné aux adultes. Après un mois d'essai, le sirop est mis en magasin dans les villes en test pour mesurer l'achat.

Au moment du recrutement, un questionnaire est administré en face à face avec la personne chargée des achats dans les foyers sélectionnés : il porte sur les habitudes d'achats et de consommation des sirops de fruit. L'image de leur marque habituelle est évaluée avec un questionnaire composé de 13 items (sous la forme d'une échelle de Likert en 7 points). Au terme de l'essai (trois semaines), ce même questionnaire est administré, toujours en face à face, pour évaluer la perception de la marque testée. Cette dernière a été confrontée ainsi systématiquement aux quatre autres marques nationales (deux marques testées par foyer : la marque habituelle et celle en test). Chaque paire est testée auprès de 60 foyers et un total de 452 observations sont exploitables après agrégation des données.

L'objectif est de vérifier si, après essai et sous l'effet de la campagne publicitaire, le positionnement voulu pour cette nouvelle marque a une réelle valeur concurrentielle face aux concurrents nationaux. Le contenu des items est le suivant : degré de concentration et qualité perçue du sirop, teneur en fruits et diversité de l'offre en parfums, restitution du goût de fruit, naturalité, niveau de prix, taux de sucre et caractère désaltérant, confiance accordée à un sirop plutôt destiné aux adultes et qui peut être servi à ses amis. Par ailleurs, dans la perspective d'une éventuelle campagne nationale en cas de lancement, l'agence de publicité a imposé un item pour évaluer si le style de sa campagne était bien perçu comme elle le souhaitait : « marque qui fait une publicité sympathique et gaie ».

1 Vérifier les conditions d'application d'une analyse factorielle métrique

Pour appliquer une analyse factorielle, il convient de respecter quelques conditions préalables, même si elles sont peu contraignantes :

– Il est recommandé de disposer d'au moins cinq variables qui, impérativement, doivent être métriques (échelle d'intervalles ou échelle de rapport) ; néanmoins, certaines variables ordinales, telles les échelles de Likert considérées comme des variables de quasi-intervalles et les variables dichotomiques peuvent être utilisées.
– Il est souhaitable que le nombre d'observations soit au moins dix fois supérieur au nombre de variables.
– Il faut un nombre suffisant de variables corrélées. Cependant, si toutes les variables sont très fortement corrélées entre elles ($r \approx 1$), un seul facteur les regroupera toutes ; si aucune variable n'est corrélée, chacune étant indépendante ($r \approx 0$), il y aura autant de facteur que de variables initiales ; dans ces deux conditions, l'usage d'une analyse factorielle n'a aucune raison d'être.

1.1 Contrôler la pertinence des variables

Sélectionner des variables pertinentes

Il faut s'assurer que les variables introduites dans l'analyse sont toutes pertinentes en regard du problème à étudier et s'assurer qu'elles sont toutes cohérentes les unes avec les autres. Leur choix repose sur une bonne maîtrise du champ étudié ou sur des hypothèses issues d'informations ou d'analyses préalables.

> Dans le cas « Sirop de fruits », alors que tous les items concernaient la perception du produit, l'agence de communication a imposé un item sur la qualité de sa campagne publicitaire : « c'est une publicité sympathique et gaie ». Comme cet item relève d'un contenu différent de l'ensemble des autres, sa pertinence pour étudier le positionnement est discutable.

Il faut vérifier que *toutes* les données nécessaires à l'analyse du problème ont bien été sélectionnées (*principe d'exhaustivité*). L'analyse factorielle ne traite que l'information introduite dans l'étude et ne peut combler d'éventuelles lacunes concernant la sélection des variables. La valeur opérationnelle du modèle de structuration obtenu vaudra ce que vaut l'information utilisée.

Disposer de variables statistiquement homogènes

Toutes les variables doivent être métriques et de longueur comparable. Si les variables n'ont pas la même dimension, les plus longues génèrent naturellement plus de variance, leur donnant artificiellement plus de poids : il faut alors travailler sur des variables standardisées (moyenne : 0 ; écart-type : 1). Les algorithmes d'analyse factorielle proposent cette transformation (généralement « par défaut »).

© Dunod – Toute reproduction non autorisée est un délit.

Dans le cas « Sirop de fruits », toutes les variables sont de même longueur puisqu'elles correspondent à une échelle de Likert en 7 points (quasi échelle d'intervalles) : elles ont donc toutes le même poids.

Si on souhaite comparer des téléphones sur la base d'un ensemble de variables objectives, telles que l'autonomie de communication et de veille, le volume, le poids et le prix, on constate que la longueur de leurs échelles est très variée et différente (tableau 4.1).

Après transformation des variables initiales en données centrées réduites, toutes les variables sont de moyenne nulle (0) et d'écart-type unitaire (1) : chaque observation est donc située par rapport à cette moyenne et la nouvelle valeur calculée exprime, en écart-type, la distance de l'observation à la moyenne 0 de l'échantillon étudié (colonnes grisées du tableau 4.1).

Tableau 4.1 – Les caractéristiques des téléphones portables

Téléphones	PRIX (€)	*zPRIX*	VOLUME (mm^3)	*zVOLUME*	POIDS (gr)	*zPOIDS*	Autonomie veille (AUTOV)	*zAUTOV*	Autonomie communication (AUTOC)	zAUTOC
T1	159	*-1,115*	85050	*0,253*	92	*-0,989*	450	*1,807*	4,0	*1,753*
T2	299	*-0,103*	54277	*-1,867*	91	*-1,059*	348	*0,566*	3,5	*0,573*
T3	399	*0,619*	94529	*0,907*	115	*0,603*	288	*-0,163*	3,5	*0,573*
T4	89	*-1,621*	80496	*-0,060*	91	*-1,059*	300	*-0,017*	3,0	*-0,606*
T5	399	*0,619*	98980	*1,214*	125	*1,296*	300	*-0,017*	3,0	*-0,606*
T6	399	*0,619*	80370	*-0,068*	115	*0,603*	220	*-0,990*	3,0	*-0,606*
T7	449	*0,981*	75871	*-0,378*	115	*0,603*	204	*-1,185*	2,8	*-1,078*
Moyenne	313,29	*0*	81367,57	*0*	106,29	*0*	301,43	*0*	3,26	*0*
Écart-type	128,05	*1*	13431,56	*1*	13,36	*1*	76,11	*1*	0,39	*1*

1.2 Vérifier que les données sont factorisables

Disposer de variables statistiquement homogènes

L'analyse factorielle métrique est basée sur l'analyse des corrélations qui existent entre les variables. L'application du modèle ne se justifie que s'il y a un nombre suffisant de variables significativement corrélées entre elles (sinon, il y aura autant de *facteurs uniques* que de variables indépendantes initiales).

Un examen de la matrice des corrélations permet de voir le nombre de corrélations significatives et la manière dont elles se répartissent entre les variables. Plus les corrélations entre certaines variables sont élevées et plus forte est la probabilité que ces variables soient représentées par un facteur. Quoi qu'il en soit, l'analyse n'a de sens que si la matrice des corrélations contient un nombre suffisant de corrélations

dont la valeur absolue est supérieure à 0,3. Les logiciels proposent divers outils pour vérifier si la matrice des corrélations est bien factorisable.

Les différents tests pour vérifier la pertinence d'une factorisation

– *La multicolinéarité restreinte*. La multicolinéarité correspond au cas où la valeur d'une variable peut être prédite par celle d'une ou de plusieurs autres variables. Si la multicolinéarité reste modérée, cela ne pose pas de problème pour réaliser une analyse factorielle.

> **Exemple.** Les deux variables « année de naissance » et « âge » partagent cette caractéristique (appelée « singularité ») puisqu'il existe une relation linéaire parfaite entre elles : l'une peut prédire exactement l'autre : « âge » = « année en cours » – « année de naissance ».

– Le niveau de multicolinéarité peut être détecté en calculant le déterminant de la matrice de corrélations qui varie entre 0 et 1. Mais en analyse factorielle, ces deux valeurs extrêmes posent problème : la valeur 0 indique qu'il y a au moins un cas de très forte multicolinéarité parmi les variables ; et la valeur 1 : que tous les coefficients de corrélation sont nuls (dans ce cas, la matrice de corrélation est une matrice d'identité). Il est conseillé de ne pas effectuer d'analyse factorielle si le déterminant est inférieur à 0,00001. Un déterminant différent de zéro signifie que la matrice des corrélations n'est pas « singulière », c'est-à-dire qu'il n'y a pas de forte multicolinéarité.

– *Le test de sphéricité de Bartlett* permet de vérifier s'il existe des corrélations suffisamment élevées entre un certain nombre de variables pour justifier une analyse factorielle. Son principe est simple : l'hypothèse H_o postule que la matrice des corrélations est une matrice d'identité (toutes les corrélations sont nulles). Avec un test d'hypothèse nulle fondé sur le Chi-deux, le test de Bartlett vérifie que la matrice des corrélations a une structure aléatoire ; si la probabilité critique est inférieure au seuil usuel de 5 % (rejet de l'hypothèse nulle) la matrice est factorisable.

> Dans le cas « Sirop de fruits », la significativité du test de Bartlett est de 0,000 : l'hypothèse nulle (toutes les corrélations sont nulles) étant rejetée, les données sont donc *a priori* factorisables (tableau 4.2).

Tableau 4.2 – Test de Bartlett et Indice KMO appliqués sur la matrice des corrélations (cas « Sirop de fruits »)

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin (K-M-O).		.848	> .70
Test de sphéricité de Bartlett	Khi-deux approximé	4097,846	
	Degrés de liberté (ddl)	78	
	Signification de Bartlett	.000	< .05

© Dunod – Toute reproduction non autorisée est un délit.

– *L'indice de Kaiser-Meyer-Olkin (KMO)* permet de vérifier que l'ensemble des corrélations partielles entre variables (le lien spécifique entre chaque paire de variables) est faible, puisque l'extraction des facteurs repose sur la variance *commune* (et non spécifique) aux variables. Il mesure la différence d'amplitude entre les coefficients de corrélation et les corrélations partielles. On vérifie ainsi que la cohérence de l'ensemble des variables est suffisamment élevée pour constituer un ou plusieurs facteurs.
– La valeur de cet indice doit être supérieure à 0,5 et aussi grande que possible pour que la matrice soit factorisée. Les concepteurs du KMO considèrent même que sa valeur devrait être supérieure à 0,70, signe que la variabilité des données aurait bien une source commune traduite par les corrélations (pour les variables introduites dans le cas « Sirop de fruits », il est de « bonne qualité » puisqu'il vaut 0,848). Le tableau 4.3 fixe les degrés d'acceptabilité de cet indice.

Tableau 4.3 – Les degrés d'acceptabilité de l'indice KMO

Inacceptable	Très faible	Faible	Moyenne	Bonne	Très bonne
< 0,5	0,5 – 0,6	0,6 – 0,7	0,7 – 0,8	0,8 – 0,9	0,9 – 1,0

Remarque

Certains logiciels permettent de calculer deux matrices dites « anti-image ». L'une est la matrice inverse des coefficients de corrélation entre paires de variables (corrélations partielles) et l'autre, la matrice inverse des covariances entre paires de variables (covariances partielles). L'ensemble des variables retenues pour l'analyse est cohérent si la plupart des éléments hors de la diagonale sont petits. Par ailleurs, la diagonale de la matrice inverse (anti-image) des corrélations correspond aux valeurs des indices KMO. C'est une information très utile quand le KMO global est faible, car elle renseigne sur les variables obtenant un coefficient inférieur à 0,5. Le KMO global est la moyenne des coefficients situés dans la diagonale de la matrice des corrélations anti-image.

2 Choisir la méthode de factorisation et construire le modèle factoriel adapté à la problématique étudiée

2.1 Les différentes méthodes de factorisation

Réduction et structuration sont les deux fonctions complémentaires de l'analyse factorielle, mais chacune a sa finalité propre qui est à l'origine de modèles d'analyse différents : analyse factorielle en composantes principales (ACP) et analyse factorielle en facteurs communs et spécifiques (AFCS, dite aussi analyse factorielle classique). Le tableau 4.4 compare ces deux méthodes. Cette comparaison peut aider à lever la perplexité de certains utilisateurs de logiciels quand ils constatent que sous un même nom, il existe deux modèles ne donnant pas des résultats identiques (*cf.* le tableau 4.5 qui produit 3 ou 4 facteurs significatifs selon le modèle utilisé).

L'analyse en composantes principales (ACP)

C'est le modèle à privilégier quand l'objectif de l'étude est de *réduire le nombre de variables brutes* pour les remplacer par des axes *indépendants* (fonction de réduction stricte). Par conséquent, il faut autant d'axes que de variables initiales pour rendre compte de toute la variance des données.

Comme le but est de condenser les mesures initiales pour les remplacer par des mesures composites (chaque facteur est une fonction linéaire de *toutes* les variables d'origine), aucune rotation des facteurs ne se justifie (c'est pourquoi certains logiciels d'ACP ne proposent pas de rotation). *En pratique*, à défaut de retenir tous les axes comme il se devrait, l'analyste choisit un nombre d'axes suffisant pour remplacer les mesures initiales et rendre compte de manière satisfaisante d'une part importante de la variance totale (et supérieure à 50 %, au minimum).

Tableau 4.4 – Les deux méthodes d'analyse factorielle métrique

	Composantes principales	Facteurs communs et spécifiques
Finalité	Remplacer *n* variables en corrélation par *n* facteurs indépendants	Rechercher des regroupements de variables qui soient interprétables
Calculs	Maximiser la somme des carrés des corrélations entre variables et facteur (saturations)	Maximiser la somme des saturations (corrélations variables/facteurs)
Cinq traits différenciateurs majeurs	Condenser les mesures initiales pour les reconstruire (et reconstruire les corrélations *a fortiori*)	Condenser les corrélations entre les variables pour reconstruire les seules corrélations
	La diagonale de la matrice des corrélations contient la variance totale propre à chaque variable (ou 1 si les variables sont standardisées)	La diagonale de la matrice des corrélations contient la communauté des variables, laquelle est estimée
	Rendre compte de la totalité de la variance avec autant de facteurs que de variables initiales	Rendre compte correctement des corrélations avec le moins de facteurs possibles
	Pour chaque facteur, rendre maximum la somme des carrés des saturations qui représente la variance expliquée	Pour chaque facteur, rendre maximum la somme des coefficients de corrélation
	Toute rotation est inutile, l'objectif étant de condenser statistiquement les mesures initiales	La rotation est une nécessité, l'objectif étant de rechercher une structure (plus ou moins latente) interprétable

L'analyse en facteurs communs et spécifiques (AFCS) ou analyse factorielle classique

C'est le modèle à privilégier si l'objectif est de *rechercher des regroupements de variables* sur la base des corrélations (les *facteurs communs*) pour mettre à jour une structure latente issue de ces regroupements qui soit facilement interprétable.

© Dunod – Toute reproduction non autorisée est un délit.

Quelques *exemples d'usages* :
– identifier les structures d'attitudes qui régissent le comportement des clients (comme les facettes de l'implication, les systèmes d'attitudes déterminant des styles de vie, etc.) ;
– élaborer le système de représentation concurrentiel ou le positionnement perçu par les prospects sur la base de variables pertinentes pour caractériser les produits ou les marques d'un marché concurrentiel ;
– mettre en évidence la structure implicite d'une batterie d'items bruts (issus d'une série d'entretiens par exemple) pour valider leur pertinence par rapport au problème étudié et, accessoirement, supprimer l'information inutile ;
– construire des batteries d'items en conformité avec des « construits » qui ont été préalablement validés.

Un facteur est un groupe de variables inter-corrélées qui forme une structure indépendante (orthogonalité des axes). *En pratique*, l'analyste recherche le plus petit nombre possible de facteurs nécessaires pour rendre compte de la seule corrélation entre les variables. Il essaie aussi de déterminer quel est le contenu commun qui, dans un même facteur, lie un ensemble de variables. Pour ce faire, une rotation des axes est souvent nécessaire pour faciliter l'interprétation des structures latentes dont les variables ne sont que la manifestation.

Le modèle en « facteurs communs » s'appuie sur la variance commune – la « communauté » (« *communality* » en anglais) – qui existe entre les différents groupes de variables dont chaque facteur rend spécifiquement compte.

Il existe plusieurs méthodes d'extraction des facteurs : par exemple, SPSS en propose sept, complétées par cinq méthodes de rotation. *Dans la pratique*, les algorithmes proposés sont une application de l'analyse des composantes principales qui traite la matrice des corrélations[1] et c'est le contenu de la diagonale de cette matrice qui précise le type de modèle utilisé :

– si *la diagonale contient l'unité* (cas des variables standardisées, de moyenne 0 et d'écart-type 1 et c'est souvent la solution par défaut dans les logiciels) : le résultat est celui d'une analyse en composantes principales, avec possible rotation des facteurs extraits (c'est ce dont nous traiterons dans ce chapitre) ;
– si *la diagonale contient la communauté estimée* des variables, c'est-à-dire la fraction de la variance totale partagée par toutes les variables introduites dans l'analyse : le résultat est celui d'une analyse en facteurs communs, avec les rotations et l'analyse spécifique des communautés par variable.

1. Dans les analyses exploratoires qui caractérisent nombre des sujets traités en gestion, les variables sont standardisées et c'est la matrice des corrélations qui est le plus souvent traitée. Il est plus rare d'utiliser la matrice des covariances (mesure des variations entre des variables seulement centrées). Comme le même algorithme appliqué aux deux matrices ne fournit pas les mêmes résultats, il faut donc justifier son choix. Cependant, dans les analyses factorielles confirmatoires, c'est la matrice de covariance qui est systématiquement utilisée.

Traitées selon l'une ou l'autre méthode, les données du cas « Sirop de fruits » ne donnent pas les mêmes résultats (tableau 4.5) : avec l'ACP, 4 facteurs significatifs rendent compte de 81,83 % de l'inertie initiale, alors que 3 facteurs significatifs suffisent avec l'AFCS pour « expliquer » près des 100 % de la variance totale.

Tableau 4.5 – Résultats selon les deux méthodes d'analyse factorielle métrique

A - Analyse en composantes principales				B - Analyse en facteurs communs et spécifiques			
Numéro composante	Valeurs propres	Pourcentage de variance	Pourcentage cumulé	Numéro facteur	Valeurs propres	Pourcentage de variance	Pourcentage cumulé
1	4,00973	30,844	30,844	1	3,77225	41,828	41,828
2	3,23922	24,917	55,761	2	2,99848	33,248	75,076
3	2,38693	18,361	74,122	3	2,09917	23,276	98,352
4	1,00208	7,708	81,830	4	0,06175	0,685	99,037
5	0,38751	2,981	84,811	5	0,04155	0,461	99,497
6	0,37769	2,905	87,717	6	0,03492	0,387	99,885
7	0,34750	2,673	90,390	7	0,00955	0,106	99,991
8	0,29581	2,276	92,665	8	0,00085	0,009	100,000
9	0,23707	1,824	94,489	9	0,0	0,000	100,000
10	0,23084	1,776	96,265	10	0,0	0,000	100,000
11	0,20506	1,577	97,842	11	0,0	0,000	100,000
12	0,15599	1,200	99,042	12	0,0	0,000	100,000
13	0,12453	0,958	100,000	13	0,0	0,000	100,000
Total	*13*	*100*	-		*9,01852*	*100*	-

Pour aller plus loin – Quelques éléments sur les principes de l'extraction des facteurs en ACP

Les données soumises à l'ACP sont sous forme d'une matrice rectangulaire de n lignes et de p colonnes (n est le nombre des individus statistiques et p le nombre des variables initiales quantitatives). Statistiquement parlant, il s'agit alors de deux espaces vectoriels qui sont considérés différemment : espace lignes et espace colonnes. Le terme χ_{ij} représente alors la valeur prise par l'individu i ($1 \leq i \leq n$) pour la variable j ($1 \leq j \leq p$).

© Dunod – Toute reproduction non autorisée est un délit.

☞

Étude des individus statistiques (les observations)

Chaque ligne représente le profil multidimensionnel d'un individu statistique (selon les modalités prises par les *p* variables initiales). On obtient aussi un individu statistique moyen en calculant les moyennes des p variables. Un individu peut alors être représenté par un espace de p dimensions. En considérant l'ensemble des individus, on obtient un nuage de n points multidimensionnel où chaque point représente un individu. Le centre de gravité de ce nuage est alors le point *G* qui représente l'individu moyen. Pour analyser la ressemblance entre deux individus *i* et *k*, l'ACP utilise la distance euclidienne donnée par la formule :

$$d^2(i,k) = \sum_{j=1}^{p} \left(x_{ij} - x_{kj}\right)^2$$

L'objectif de l'analyse factorielle est de résumer les *n* profils par des variables synthétiques dont le nombre est inférieur à *n*. Dans un premier temps, l'ACP procède par un centrage et une réduction des données.

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}$$

En termes d'individus, l'opération de centrage consiste à placer l'origine au centre de gravité du nuage alors que la réduction fait en sorte que le poids de chaque variable devienne identique. Géométriquement parlant (ce ne peut être représenté graphiquement pour des espaces de plus de trois dimensions), chaque point du nuage peut alors être représenté par un vecteur de *p* dimensions ayant son point initial à l'origine, le centre de gravité du nuage.

Concrètement, l'ACP cherche une première représentation axiale des points du nuage, c'est-à-dire une droite (axe u_1) qui passe par le centre de gravité. L'équation de cette première droite est obtenue de sorte que la somme des projections orthogonales des vecteurs (points du nuage) soit maximale. Si pour le point M_i du nuage, la projection orthogonale du vecteur OM_i sur l'axe u_1 est notée OH_i, la somme des carrés des projections de l'ensemble des *n* points est déterminée par l'expression

$$\sum_{i=1}^{n} w_i \left(OH_i\right)^2$$

où les coefficients représentent le poids de chacun des points du nuage de sorte que $\sum_{i=1}^{n} w_i = 1$. L'ACP attribue des poids égaux (1/*n*) à tous les points du nuage. En d'autres termes, lors de l'analyse des individus, l'ACP cherche à maximiser la somme :

$$\frac{1}{n}\sum_{i=1}^{n} \left(OH_i\right)^2$$

L'inertie étant définie comme le rapport entre la masse et le carré des distances, maximiser la somme des carrés de projections orthogonales revient alors à maximiser l'inertie.

L'étape suivante de l'ACP consiste à chercher une deuxième droite u_2 qui passe toujours par le centre de gravité des points, mais avec la contrainte d'être orthogonale avec le

☞

☞

premier axe en plus de celle de maximiser la somme des projections des points sur elle. Les deux premières droites (axes) définissent ainsi le premier plan factoriel sur lequel l'ensemble des *n* points est projeté, mais de telle sorte que leur projection orthogonale sur ce plan soit le moins possible déformée. L'ACP recherche ensuite un troisième axe d'inertie maximale, orthogonal aux deux premiers. Sous ces conditions, l'ACP détermine ainsi autant d'axes orthogonaux entre eux qu'il y a des variables initiales. Chaque nouvel axe doit être orthogonal avec les axes précédents. Le premier axe u_1 est l'axe qui représente le mieux la variabilité des données (« explique » le plus de variance), l'axe u_2 celui qui explique le mieux la part de variance restante, et ainsi de suite, avec la contrainte que les *p* axes (u_1, u_2, … u_p) rendent compte de la totalité des variations de l'échantillon.

La détermination des axes d'inertie maximum est obtenue par le calcul matriciel (dont la description dépasse le cadre de cet ouvrage). Sur le plan mathématique, les *p* axes d'inertie maximale ne sont rien d'autres que les vecteurs propres de la matrice des corrélations entre les variables alors que les sommes des projections des points sur chacun des axes sont les valeurs propres. L'ACP calcule les coordonnées des individus sur chacun des axes ce qui permet de visualiser la projection des individus sur les plans factoriels. Les axes factoriels étant des variables synthétiques fixant les coordonnées de chaque individu, l'ACP rend possible le calcul des coefficients de corrélations entre les axes factoriels et les variables initiales, ainsi que la représentation des *p* variables sur les plans factoriels avec pour coordonnées, ces coefficients de corrélation.

Étude des variables

Lors de l'analyse des variables (colonnes), il n'est pas censé d'analyser la ressemblance entre variables qui peuvent présenter des caractères très différents, par exemple la taille et le poids d'un individu. L'ACP analyse plutôt la liaison entre les variables, laquelle est mesurée par les coefficients de corrélation linéaire. Il s'agit alors de considérer un nouvel espace de *p* vecteurs de n dimensions chacun. Cet espace définit un nuage constitué de n points qui peuvent être représentés par des vecteurs dont l'origine, contrairement à la vectrice ligne, est bien zéro. Pour analyser la proximité entre deux variables, l'ACP mesure l'angle, ou plus exactement le *cosinus* de l'angle, entre ces deux variables ; et si les données sont centrées réduites, cette mesure n'est rien d'autre que le coefficient de corrélation entre les deux variables. La longueur d'une variable centrée réduite (vecteur colonne) étant égale à 1, tous les points du nuage se trouvent alors sur la sphère de rayon 1 et de centre 0.

Dans l'étude de l'espace des variables, l'objectif de l'ACP est de visualiser le nuage des points sur des plans factoriels où ces points ont été projetés, mais de façon à déformer le moins possible ce nuage. De manière similaire à l'étude des lignes, l'ACP cherche une suite de p axes orthogonaux d'inertie maximale. Le premier axe est tel que la somme des carrés des coefficients de corrélation de l'ensemble des *n* variables initiales pour cet axe soit maximale, le deuxième un peu moins et ainsi de suite. L'ACP permet alors de représenter les p variables sur les plans factoriels. Cette projection est exactement identique à celle obtenue lors de l'étude des individus, mais l'interprétation est différente.

© Dunod – Toute reproduction non autorisée est un délit.

2.2 Décider du nombre de facteurs à retenir dans le modèle

Le modèle factoriel construit par l'analyste repose sur le choix du nombre de facteurs fondé sur les valeurs propres.

Signification de la « valeur propre »

La *valeur propre* (« *eigenvalue* ») ou *contribution* – symbolisée par la lettre λ – (lambda). Statistiquement parlant, la valeur propre liée à un facteur donné représente la somme des carrés des coefficients de corrélation des variables initiales avec ce facteur. Cette somme étant égale à la somme des coefficients de détermination, elle représente alors la part de la variance de l'ensemble des variables initiales expliquée par le facteur. Chaque valeur propre est donc associée à un facteur qui rend compte d'une fraction de la variance totale générée par les données initiales. La variance totale correspond à la somme de la variance des variables d'origine qui sont reportées dans la diagonale de la matrice des corrélations :

– **Dans l'analyse en composantes principales** où les variables sont standardisées (moyenne 0 et variance 1), la variance totale est donc égale au nombre de variables introduites dans l'analyse (la diagonale n'est composée que de 1).

> Dans le cas « Sirop de fruits » (tableau 4.5 A), la variance totale – ou somme des valeurs propres – vaut 13 puisque l'analyse inclut 13 items dont les mesures ont été standardisées.

– **Dans l'analyse en facteurs communs et spécifiques**, la diagonale contient la valeur estimée des communautés initiales ; elles correspondent au carré des coefficients de corrélation multiple calculé entre chaque variable et l'ensemble des autres variables introduites dans l'analyse.

> Pour le même cas (tableau 4.5 B), la valeur totale des communautés estimées relative aux 13 items vaut 9,01852.

Méthodes pour déterminer les facteurs significatifs

Quel que soit le modèle d'extraction des facteurs utilisé (composantes principales ou facteurs communs), plusieurs méthodes permettent de déterminer le nombre de facteurs « utiles » pour rendre compte du maximum de la variance totale.

Sélectionner le nombre de facteurs significatifs

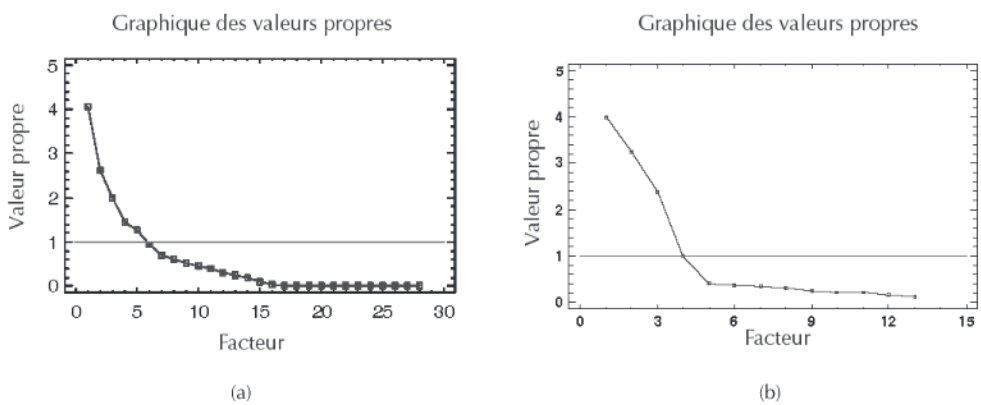
– Les *variables sont standardisées* (variance = 1). Pour prétendre condenser une partie de l'information initiale, un facteur doit prendre en charge une quantité de variance au moins supérieure à celle d'une variable. Donc, seuls les facteurs dont la valeur propre est supérieure à 1 sont significatifs : c'est le critère de choix de Kaiser où $\lambda > 1$.
– Les *variables ne sont pas standardisées*. L'histogramme de la distribution des valeurs propres (en ordonnée) en fonction de l'ordre décroissant des facteurs (en abscisse) a la forme d'une courbe. L'accentuation de la courbe marque un seuil qui

définit le nombre de facteurs à retenir : on retient les axes qui précèdent ce seuil, car la contribution à la variance expliquée des axes suivants devient marginale. C'est le test proposé par Cattell, connu sous le nom de « test des éboulis » (« scree test ») ou « test du coude » (« elbow test »).

La figure 4.1 illustre cette démarche. En A, l'analyse porte sur 27 variables ; un premier coude est visible entre le 4e et le 5e facteur, puis un second entre le 6e et le 7e avant de constater un éboulis à partir du 15e facteur. Il faut donc choisir un modèle en 4 ou 6 facteurs. En B, un seul coude est visible entre les facteurs 4 et 5 : il faut donc retenir le modèle en 4 facteurs.

De manière pratique, on peut aussi considérer le pourcentage de variance expliquée en fonction du nombre de variables initiales. Si on fait l'hypothèse qu'aucune d'elles ne privilégie un quelconque facteur, alors on peut admettre que tout facteur s'approchant du ratio 100/*n* (où *n* est le nombre de variables) peut être retenu, car il prend en charge une quantité de variance au moins égale à celle d'une variable (c'est l'approche rencontrée dans l'analyse des correspondances simples – *cf.* chapitre 3).

Dans l'exemple de la figure 4.1-A (cas « train »), ce pourcentage minimum est de 100/27 = 3,70 % ; et dans le cas « Sirop de fruits » (figure 4.1-B), de 100/13 = 7,69 % : les facteurs retenus sont supérieurs à cette valeur seuil (7,71 % pour le 4ième facteur de l'ACP – *cf.* tableau 4.6).



Figures 4.1 – Histogramme des valeurs propres ou « scree test »

(A- 27 variables initiales du cas « train » ; B - 13 variables du cas « Sirop de fruits »)

Évaluer la contribution des facteurs

Il faut ensuite considérer la proportion de variance expliquée par les facteurs répondant aux critères de sélection ci-dessus (les facteurs « utiles » pour prendre en charge la variance commune). En toute logique, tout modèle retenu doit pouvoir rendre compte de plus de la moitié de la variance totale (sinon, la variance spécifique aux variables est plus importante que leur variance commune). Il est souhaitable que

© Dunod – Toute reproduction non autorisée est un délit.

le pourcentage de variance expliquée soit le plus élevé possible (un seuil minimal de 60 % correspond à un modèle acceptable ; cependant les objectifs d'étude peuvent parfois conduire à se contenter d'un seuil inférieur).

> Dans le cas « Sirop de fruits », la somme des valeurs propres (première ligne du tableau 4.6), soit la variance commune à l'ensemble des variables, vaut 10,638 (= 4,010 + 3,329 + 2,387 + 1,002). Cette valeur définit la quantité totale de variance expliquée par le modèle en quatre facteurs ; laquelle représente 81,83 % de la variance totale (10,638/13).
>
> La variance expliquée par le modèle étant considérée comme référence (10,638 = 100 %), le ratio [contribution d'un facteur/variance expliquée totale] permet donc de déterminer le poids ou l'importance de chaque facteur dans la prise en charge (avant rotation) de la variance commune (ligne a) et totale (ligne b du tableau 4.6). Ainsi, la première composante prend en charge à elle seule plus du tiers de la variance expliquée (37,69 %), et la seconde un peu moins (30,46 %). Ensemble, ces deux facteurs représentent les 2/3 de la variance expliquée par le modèle et plus de la moitié de la variance totale (30,84 + 24,92 = 55,76 %). Comme cette valeur reste inférieure à 60 %, il faut donc considérer les autres facteurs qui sont significatifs (valeur propre supérieure à 1). C'est l'analyse du contenu des facteurs qui indiquera s'il faut retenir le modèle complet à 4 facteurs (près de 90 % de la variance totale étant « expliquée » ou si le modèle à 3 facteurs suffit (74 %).

Tableau 4.6 – Tableau des résultats d'une analyse factorielle (« Sirop de fruits »)

	RÉSULTATS AVANT ROTATION				
FACTEURS	1	2	3	4	Total
Valeurs propres (λ) :	4,010	3,329	2,387	1,002	10,638
a : % variance expliquée	37,690	30,460	22,430	9,420	100,00
b : % variance totale	30,840	24,920	18,360	7,710	81,830

3 Valider la pertinence et interpréter les résultats du modèle retenu

Le calcul statistique définit les axes qui résument l'information initiale en rendant compte du maximum de la variance commune aux données traitées. L'interprétation de ces axes repose alors sur une analyse de la matrice des corrélations variables-facteurs ou *matrice des saturations*. Mais encore faut-il que ce contenu soit interprétable : une éventuelle rotation des axes – dont l'objectif est d'optimiser la relation entre les variables initiales et chacun des facteurs – peut faciliter cette tâche.

3.1 Décider d'effectuer une rotation

Compte tenu du premier objectif de l'analyse factorielle qui est de condenser l'information, il se peut que la première solution mathématique proposée ne permette pas d'interpréter facilement le contenu des facteurs (ou la logique des

structures latentes mises à jour si c'est l'objectif de l'étude). Par exemple, il n'est pas rare qu'une majorité des variables initiales soit liée avec le premier facteur. Effectuer une rotation des axes consiste à rechercher de manière assez empirique, une autre répartition des saturations qui rend les facteurs plus facilement interprétables car correspondant à une « structure simple » (Thurstone).

Remarque

Une analogie permettra de comprendre la mécanique de l'analyse factorielle et l'intérêt de la rotation. Lors d'un déménagement, on souhaite gagner le maximum de volume pour minimiser le nombre de caisses (fonction de réduction). Ranger les livres en fonction de la seule place disponible dans les caisses peut rendre très difficile la récupération aisée d'un ouvrage. Mais si les ouvrages sont rangés par format et, à format équivalent, par collection, la récupération devient moins aléatoire (fonction de structuration). La « rotation » correspond en quelque sorte à la recherche d'un autre système de classement qui respecte cependant la contrainte de volume.

Tableau 4.7 A – Distribution des saturations et des valeurs propres avant rotation

A - Matrice des corrélations item/facteurs (« saturations ») avant rotation					
	Facteur 1	Facteur 2	Facteur 3	Facteur 4	*Communauté*
Item 1 *bon goût de fruit*	0,9402	-0,0972	-0,0176	0,0177	*0,8942*
Item 2 *naturel*	0,9157	-0,1203	0,0004	-0,0045	*0,8530*
Item 3 *variété des parfums*	0,8721	-0,1427	0,0109	-0,0001	*0,7811*
Item 4 *riche en fruit*	0,1559	0,9101	-0,1325	-0,0194	*0,8705*
Item 5 *publicité gaie*	0,0179	-0,0150	0,0429	0,9982	*0,9989*
Item 6 *pour adultes*	0,0267	0,1237	0,8874	-0,0332	*0,8046*
Item 7 *avec des amis*	0,0196	0,1246	0,9072	0,0215	*0,8395*
Item 8 *concentration*	0,0982	0,9088	-0,1040	0,0225	*0,8469*
Item 9 *désaltérant*	-0,1647	-0,8260	0,1038	-0,0096	*0,7212*
Item 10 économique	-0,8645	0,0900	0,0022	0,0346	0,7567
Item 11 *niveau de sucre*	0,1559	0,8798	-0,0682	0,0106	*0,8032*
Item 12 *avec confiance*	-0,0689	-0,1554	-0,8535	0,0365	*0,7587*
Item 13 *haute qualité*	0,8290	-0,1408	-0,0408	0,0002	*0,7088*
Valeur propre ou **contribution**	**4,0092**	**3,2379**	**2,3867**	**1,0019**	**10,638**

Il existe plusieurs techniques de rotation. La plus courante est la rotation Varimax (pour « *var*iances *max*imisées »). Cette rotation orthogonale des facteurs (qui préserve leur indépendance) consiste à *répartir de manière différente la même variance commune initiale* entre les variables et les facteurs, c'est-à-dire la « communauté » (qui vaut 10,638 dans le cas traité, tableaux 4.6 et 4.7). La rotation consiste à maximiser dans chaque facteur l'écart entre les saturations les plus élevées (proche de 1) et minimiser les saturations les plus faibles (proche de 0).

© Dunod – Toute reproduction non autorisée est un délit.

Tableau 4.7 B – Distribution des saturations et des valeurs propres après rotation

B - Matrice des corrélations item/facteurs (« saturations ») après rotation Varimax					
	Facteur 1	Facteur 2	Facteur 3	Facteur 4	*Communauté*
Item 1 *bon goût de fruit*	**0,9442**	*0,0433*	*0,0013*	*0,0257*	*0,8942*
Item 2 *naturel*	**0,9233**	*0,0145*	*0,0160*	*0,0039*	*0,8530*
Item 3 *variété des parfums*	**0,8833**	*-0,0151*	*0,0218*	*0,0085*	*0,7812*
Item 10 économique	**-0,8685**	*-0,0368*	*-0,0162*	*0,0268*	*0,7568*
Item 13 *haute qualité*	**0,8412**	*-0,0125*	*-0,0306*	*0,0071*	*0,7088*
Item 4 *riche en fruit*	*0,0216*	**0,9323**	*-0,0074*	*-0,0291*	*0,8705*
Item 8 *concentration*	*-0,0359*	**0,9193**	*0,0175*	*0,0131*	*0,8469*
Item 11 *niveau de sucre*	*0,0249*	**0,8944**	*0,0514*	*0,0028*	*0,8033*
Item 9 *désaltérant*	*-0,0420*	**-0,8481**	*-0,0096*	*-0,0014*	*0,7212*
Item 6 *pour adultes*	*-0,0046*	*0,0068*	**0,8969**	*-0,0102*	*0,8047*
Item 7 *avec des amis*	*-0,0125*	*0,0046*	**0,9150**	*0,0450*	*0,8396*
Item 12 *avec confiance*	*-0,0329*	*-0,0485*	**-0,8689**	*0,0143*	*0,7587*
Item 5 *publicité gaie*	*0,0104*	*-0,0074*	*0,0154*	**0,9992**	*0,9989*
Valeur propre ou **contribution**	**3,9916**	**3,2398**	**2,4019**	**1,0033**	**10,638**
% de variance expliquée	**37,522**	**30,455**	**22,578**	**9,431**	**100 %**

Dans le cas « Sirop de fruits », la comparaison des matrices des coefficients de saturation avant et après rotation (tableau 4.7-A/B) montre une légère amélioration des saturations les plus fortes après rotation, et une diminution très sensible des saturations les plus faibles. Par exemple, on voit que cette diminution est très nette pour les items 4, 8, 9 et 11 du facteur 1 et pour toutes les petites corrélations du facteur 2.

Comme dans ce cas les données initiales sont fortement structurées (résultat d'une étude préalable pour déterminer les caractéristiques significatives du positionnement), on retrouve les mêmes items dans les mêmes facteurs avant et après rotation. Mais dans nombre de situations où les données ne sont pas aussi bien structurées *a priori*, on peut constater des différences de regroupement entre les variables qui migrent après rotation dans des facteurs différents, ce qui les rend plus facilement interprétables.

Avant et après rotation, la communauté reste identique – c'est une contrainte dans la recherche d'une solution par rotation orthogonale – mais la contribution de chaque facteur peut être répartie de manière différente, voire rééquilibrée entre les facteurs[1]. Dans le cas traité, les modifications sont minimes, mais elles existent bien.

1. Certains auteurs suggèrent d'effectuer la rotation orthogonale avec un facteur supplémentaire par rapport au modèle initial. Si la structure dégagée est robuste, aucune variable ne devrait être liée significativement (forte saturation) à cette composante supplémentaire. A. Jolibert et P. Jourdan : *Marketing Research. Méthodes de recherche et d'études en marketing*, Dunod, 2006, p. 313.

Focus

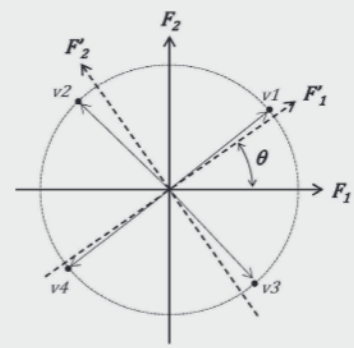
Les différentes méthodes de rotation

L'objectif de la rotation est d'obtenir des axes factoriels dont les corrélations sont très élevées avec certaines variables initiales et faibles avec les autres variables. L'ACP calcule autant d'axes que de variables initiales en maximisant l'inertie, il est donc impossible d'appliquer une rotation sur l'ensemble des *p* facteurs (problème de degrés de liberté), mais seulement sur une sélection des facteurs (usuellement, ceux qui sont significatifs).

Une rotation des axes peut s'effectuer dans n'importe quel sens et sans modifier la position relative des points du nuage. Elle peut être orthogonale ou oblique. Une rotation orthogonale préserve l'indépendance des facteurs et supprime donc les risques de multicolinéarité entre les variables. Cela permet d'utiliser ses résultats dans d'autres techniques statistiques, par exemple pour effectuer une typologie, ou dans les modèles régressifs. *Varimax*, *Quartimax* et *Equimax* sont les trois principales méthodes de rotation orthogonales ; *Oblimin* étant la rotation oblique.

La rotation Varimax. Elle répond au besoin d'expliquer les facteurs. Elle consiste à optimiser les coefficients de saturation dans les facteurs pour obtenir des coefficients proches de 1 (lien maximal d'une variable avec un facteur) ou proche de 0 (absence de lien ou indépendance avec un facteur). C'est la méthode la plus efficace pour faciliter l'interprétation des facteurs dans la mesure où elle fournit un petit nombre de variables bien saturées dans chaque facteur.

Le schéma ci-dessous présente la projection de 4 variables initiales fictives (v1, v2, v3 et v4) sur le plan factoriel F1-F2 issu de l'ACP. Une telle projection de ces variables sur le plan factoriel F1-F2 ne permet pas de les regrouper de façon pertinente parce que chacune d'elle est plus ou moins corrélée avec les deux facteurs. La rotation *varimax* consiste alors à faire tourner les facteurs de façon à ce que les variables soient corrélées le plus possible avec un des facteurs et le moins possible avec l'autre. On obtient ainsi un autre plan factoriel F'1-F'2 beaucoup plus facile à interpréter : le facteur F'1 est corrélé fortement avec les variables v1 et v4 alors que facteur F'2 l'est avec les variables v2 et v3. La rotation *varimax* diffère de l'ACP en ce qu'elle maximise la variance spécifique à chaque facteur et non seulement la variance expliquée par le premier facteur.



La rotation Quartimax. Elle répond au besoin d'expliquer économiquement la variance des variables. La méthode vise à réduire le nombre de facteurs nécessaires pour rendre compte de la variance de chaque variable. Une conséquence peut être la concentration des saturations les plus élevées sur un nombre très réduit de facteurs, voire un seul.

☞

© Dunod – Toute reproduction non autorisée est un délit.

☞

La rotation Equimax. Elle résulte d'un compromis entre les deux précédentes.

La rotation Oblimin est la plus connue des rotations obliques, rotations qui ne supposent pas *a priori* que les facteurs soient orthogonaux et donc indépendants. C'est une méthode adaptée à l'étude de structures latentes ou, sur le plan théorique, à la recherche d'un facteur d'ordre supérieur (telles que la personnalité ou l'intelligence qui, en psychométrie, dépendent d'un certain nombre de facettes préalablement déterminées). Sauf cas spécifiques, la méthode est peu utilisée en gestion en raison de la difficulté à interpréter ses résultats. Cependant, elle peut être utilisée dans certaines conditions pour vérifier la stabilité d'une structure factorielle orthogonale.

3.2 Savoir interpréter et utiliser les facteurs

Comprendre les indicateurs statistiques

L'interprétation des facteurs s'appuie sur la matrice des saturations après rotation si elle a eu lieu, à l'exemple du tableau 4.7-B. Commençons par préciser le vocabulaire.

Définitions

Saturation (« *loading* ») : corrélation entre une variable et un facteur dont elle mesure l'intensité du lien. L'analyse de la matrice des saturations permet d'interpréter les facteurs.

Communauté (« *communality* ») : étant le carré du coefficient de saturation (ou coefficient de détermination), elle mesure la proportion de la variance de chaque variable initiale prise en charge par chacun des facteurs identifiés. Elle permet de valider la capacité du modèle factoriel retenu à expliquer la variance de l'ensemble des variables.

La somme des communautés pour l'ensemble des variables correspond à la variance totale, soit la contribution λ totale. De ces définitions, on en déduit que :

– *Pour un facteur*, la somme (en colonne) des communautés pour l'ensemble des variables correspond à la contribution ou valeur propre du facteur, c'est-à-dire la part de la variance totale que ce facteur prend en charge.

Soit **l'exemple « Sirop de fruits »** contenu dans le tableau 4.7-B,

– sur *le facteur 1*, la somme du carré des coefficients de saturation des 13 variables vaut 3,9916 ($0{,}9442^2 + 0{,}9233^2 + \ldots + 0{,}0329^2 + 0{,}0104^2$) : cette valeur correspond à la valeur propre de ce facteur *après* rotation (elle valait 4,009 avant rotation) ;

– *pour le modèle avec quatre facteurs*, la somme des valeurs propres de chaque facteur, soit la communauté, vaut 10,638 (dernière ligne du tableau, colonne communauté). Ainsi, ce modèle rend compte de 81,83 % de la variance totale (10,638/13).

– *Pour une variable*, la somme (en ligne) du carré des coefficients de saturation pour l'ensemble des facteurs correspond à sa communauté, c'est-à-dire la part de

variance de cette variable prise en charge par le modèle factoriel. Si l'ensemble des facteurs est retenu, la communauté correspond à la variance totale de la variable (soit 1 si elle est standardisée).

Pour un nombre plus réduit de facteurs, la communauté mesure la quantité de variance de la variable prise en charge par le modèle choisi (ou sa proportion si les données sont standardisées). Si cette valeur est inférieure à 50 %, cela signifie que la variance spécifique est plus importante que la variance que cette variable partage avec l'ensemble des autres : elle n'aurait donc pas sa place dans l'analyse.

À lui seul, le facteur 1 rend compte de 89,15 % de la variance de l'item 1 ($0{,}9442^2$) et l'ensemble des quatre facteurs retenus rend compte de 89,42 % de sa variance totale ($0{,}9442^2 + 0{,}0433^2 + 0{,}0013^2 + 0{,}0257^2 = 0{,}8942$), c'est-à-dire sa communauté.

Par ailleurs, pour cinq items, le facteur 1 prend en charge la quasi-totalité (> 99 %) de leur variance prise en charge par le modèle en quatre facteurs : ce facteur 1 est donc déterminant pour rendre compte de ces cinq variables.

	Saturation dans le facteur 1	Carré de la saturation facteur 1	Communauté du modèle à 4 facteurs	% de la communauté prise en charge par Facteur 1
	(a)	(b)	(c)	b / c
Item 1 *bon goût de fruit*	0,9442	0,8915	0,8942	99,70
Item 2 *naturel*	0,9233	0,8525	0,8530	99,94
Item 3 *variété des parfums*	0,8833	0,7802	0,7812	99,87
Item 10 *économique*	- 0,8685	0,7543	0,7568	99,67
Item 13 *haute qualité*	0,8412	0,7076	0,7088	99,83

L'analyse de ces statistiques permet d'évaluer la qualité du modèle factoriel construit après sélection des facteurs :

– le ratio « Communauté (variance expliquée)/Variance totale » doit être supérieur à 50 % ; inférieur, ce ratio indique que le modèle rend mal compte de la variance *partagée* par toutes les variables ou que cette variance commune est faible, rendant donc le modèle peu opérationnel ;

■ Dans le modèle à quatre facteurs extrait, ce ratio est très élevé : 10,638/13 = 81,83 %

– la corrélation variable-facteur (le coefficient de saturation) permet de hiérarchiser les variables en fonction de leur contribution à l'émergence d'un facteur ; en conséquence, un facteur dont la contribution est forte (sa valeur propre) est un facteur lié à un nombre important de variables : il doit donc être conservé dans le modèle final ;
– un facteur qui rend compte significativement d'une variable importante en regard des objectifs de l'étude doit être conservé, sauf à décider de sacrifier cette variable.

© Dunod – Toute reproduction non autorisée est un délit.

Le facteur 4 est lié une seule variable : « publicité gaie et sympathique ». Or, on a vu que cet item 5 ajouté à la demande de l'agence de communication peut être considéré comme un « intrus » : il génère d'ailleurs un facteur « unique » qui valide cette hypothèse ; il est alors envisageable de ne pas le retenir pour garder un modèle d'évaluation des marques qui soit homogène.

Enfin, il faut se rappeler que les facteurs restituent la masse d'information contenue dans le tableau de données *du seul point de vue statistique*, c'est-à-dire celui de la variance : mais cela ne préjuge en rien, ni de l'importance, ni du sens ou de la signification des données traitées. Cependant, les relations entre variables et facteurs aident à comprendre la pertinence de la structure latente mise à jour par l'analyse et, donc, la signification des facteurs.

Interpréter les facteurs

L'interprétation du contenu de chaque facteur résulte de l'identification de toutes les variables saturées dans un même axe. L'objectif consiste à rechercher le point commun, le concept ou l'idée transversale, qui rend compte de son contenu.

Décider de l'affectation d'une variable à un facteur

Après rotation, si la structure dégagée est robuste, la quasi-totalité des variables doit être affectée à une composante et une seule. Cette affectation repose sur la valeur absolue des saturations (le signe n'intervient que dans l'interprétation) et leur significativité : pour un seuil de risque α de 5 %, le tableau 4.8 donne pour diverses tailles d'échantillon, la valeur critique du coefficient de saturation qui doit être dépassée pour être significative.

Tableau 4.8 – Valeurs significatives des saturations selon la taille d'échantillon

Taille d'échantillon	Valeur des Saturations
350	0,30
250	0,35
200	0,40
150	0,45
120	0,50
100	0,55
85	0,60
70	0,65
60	0,70
50	0,75

Traitement du cas « Sirop de fruits » (*cf.* Tableau 4.7-B)

Le **facteur 1** est celui qui prend la plus grande charge de la variance (37,5 %). Les cinq items qui le composent sont relatifs à la *perception* des qualités du produit : les « qualités

organoleptiques perçues » semblent être ce qui est commun à cet ensemble de variables. Par ailleurs, ce facteur oppose l'item « économique » (signe -) à l'ensemble des autres (signe +) : on peut en conclure que « plus c'est bon, moins c'est économique », car on en consomme d'autant plus !

Le **facteur 2** (30,5 % de la variance) associe quatre items liés aussi à ses caractéristiques, mais plutôt dans une perspective technique : leur point commun concerne ici les « caractéristiques organoleptiques objectives » d'un Sirop de fruits. Ici, l'item « désaltérant » (signe -) s'oppose à l'ensemble des autres (signe +) : « plus le sirop est fort et concentré, moins il est désaltérant ».

Alors qu'à première vue, on aurait pu penser que les neuf items liés à ces deux facteurs présentaient une certaine redondance et que leur liste aurait pu être réduite, l'analyse factorielle met en évidence une différence de point de vue dans la formulation des items : certains items relèvent du vécu, de l'expérience du produit (perception subjective dans le facteur 1), quand les autres relèvent plutôt de la composition du produit telle que l'étiquette peut la présenter (objectivation des caractéristiques dans le facteur 2).

Le **facteur 3** (22,6 % de la variance) associe trois items qui relèvent plutôt du contexte de consommation et dont le point commun pourrait correspondre au « statut social de la marque ». L'item « servir avec confiance » (signe -) s'oppose aux deux autres (signe +) : on ne sert pas à ses invités un produit dans lequel on n'a pas confiance.

Ces trois facteurs, qui constituent la base de perception des différentes marques de sirop de fruits par les consommateurs, permettent d'étudier leur positionnement avec un modèle structuré.

Le **facteur 4** (9,4 % de la variance) n'est lié qu'à un seul item qui concerne la perception de la campagne publicitaire. On s'interrogeait sur la pertinence de cet item exigé par l'agence (« fait une publicité gaie et sympathique ») : l'analyse factorielle confirme son statut particulier et l'isole dans un facteur unique. Supprimer cette variable ne modifie pas la cohérence du modèle qui reste structuré de manière identique, mais avec seulement trois facteurs.

Dénommer les facteurs

La dénomination des facteurs n'est pas contractuelle : elle relève de la responsabilité de l'analyste et non du logiciel utilisé. Elle est pourtant déterminante des représentations ou des images mentales que les auditeurs ou les lecteurs du rapport se feront de chaque facteur. Elle n'est donc pas neutre. Pour ne pas se laisser abuser par des appellations affriolantes, il est nécessaire de toujours revenir aux variables d'origine significativement liées à un facteur pour bien expliciter la signification. Cette exigence s'impose pour limiter les erreurs d'interprétation et les décisions erronées qui peuvent s'ensuivre.

Identifier et interpréter les variables « ambivalentes »

– *Cas d'une variable saturée sur plusieurs dimensions*. Une variable est dite « ambivalente » dès lors qu'elle est liée de manière significative à deux (voire plusieurs) facteurs simultanément : sa saturation oscille alors entre 0,40 et 0,60 selon les facteurs, ce qui représente au total plus de 50 % de sa variance expliquée. Ce type

© Dunod – Toute reproduction non autorisée est un délit.

de résultat, qui n'est pas exceptionnel, signifie que la variable appartient à deux concepts différents ou prend une signification différente selon le point de vue abordé. Ainsi, l'analyse factorielle permet d'élucider les différentes significations qu'un terme ou une expression peut prendre : en identifiant des termes différents mais sémantiquement redondants (car associés dans un même facteur) ou en repérant leur ambiguïté sémantique (associés simultanément à des facteurs différents). Cette analyse peut donc conduire à supprimer des variables, soit en raison de la forte redondance entre termes différents, soit en raison de leur forte ambiguïté sémantique.

– *Cas d'une variable dont la communauté est faible*. Si une variable ne présente qu'une très faible communauté tous facteurs confondus (très inférieure à 0,50), c'est alors la nature de cet item qui pose question : il relève probablement d'un champ conceptuel sans rapport avec la thématique fondamentale de l'étude. Après en avoir examiné les raisons, une telle variable doit donc être retirée de l'analyse.

Éviter les interprétations abusives

Il importe de rappeler que toute structure issue d'une analyse factorielle n'est qu'une réorganisation des données initiales fondée sur une logique statistique. Il n'y a pas de lien direct avec les aspects psychologiques auxquels ces données pourraient référer.

Il serait donc erroné d'affirmer que le facteur qui porte la plus grande valeur propre est celui qui a le plus d'importance sur un plan managérial. Plus modestement, cela signifie que ce facteur porte un grand nombre de variables parmi celles introduites dans l'étude, lesquelles renvoient à une même idée ou un même concept que ce facteur synthétise et restitue.

> Dans l'exemple des sirops de fruit, la valeur propre du premier facteur – la plus élevée des quatre facteurs (37,5 %) – indique qu'il contient le plus grand nombre de variables issues de la grille d'items utilisée pour étudier le positionnement des marques. Ce facteur met en évidence une thématique, celle des qualités perçues d'un sirop, qui est donc la thématique qui a le plus d'items dans la grille utilisée. Mais, en aucun cas, cela ne peut être considéré comme le facteur le plus important pour les prospects, ni être interprété comme les éléments auxquels les clients seraient le plus sensibles quand ils comparent ces marques.

Si, dans la matrice des données traitées, une thématique majeure a été omise par rapport à la problématique de l'étude, l'analyse factorielle ne pourra la compenser et le résultat sera erroné.

> Une société de conseil réalise une étude sur les motivations des vendeurs dans la perspective de commercialiser des plans de formation adaptés à la diversité des profils qui pourraient exister. À l'examen des items élaborés pour étudier les motivations, il est apparu qu'aucun ne faisait référence aux modes de rémunération…

La validité de l'analyse, c'est-à-dire sa conformité avec la réalité du terrain, repose sur le travail préparatoire à l'étude. Toutes les données introduites dans l'analyse doivent être justifiées en regard de la problématique traitée et s'appuyer éventuelle-

ment sur des concepts théoriques ou, à tout le moins, sur des hypothèses liant le choix des variables au problème à étudier.

Vérifier la validité des facteurs : le coefficient alpha de Cronbach

Une dernière question se pose concernant la validité des facteurs issus d'une analyse factorielle : représentent-ils bien la réalité qu'ils sont censés traduire ? On peut vouloir vérifier le degré d'homogénéité des items qui composent un facteur (particulièrement si l'ACP est utilisée pour construire une échelle), ou s'assurer qu'ils sont « équivalents » à considérer la notion qu'un facteur résume. Le coefficient de Cronbach représenté par la lettre grecque « α » est une statistique qui permet de mesurer la cohérence interne des facteurs. Il fait appel à la méthode *half-split* qui consiste à séparer aléatoirement les p items de l'échelle à construire en deux groupes contenant le même nombre d'items, puis à calculer leurs scores et à calculer ensuite la corrélation entre les deux scores partiels. L'alpha de Cronbach n'est rien d'autre que la moyenne de tous les split-half possibles. Cronbach a proposé la formule suivante qui permet de calculer approximativement cette valeur :

$$\alpha = \frac{p * \overline{r}}{1 + (p-1) * \overline{r}}$$

où p représente le nombre de variables (items) de l'échelle et $\overline{r}$ la moyenne des coefficients de corrélation de tous les couples d'items. La formule ci-dessus correspond à l'alpha standardisé : les calculs sont effectués à partir des items centrés et réduits. La valeur du coefficient de Cronbach, qui correspond à un niveau de variance partagée par les variables, varie de 0 (toutes les variables sont indépendantes) à 1 (toutes sont totalement corrélées). Une valeur proche de 1 signifie que les items mesurent bien un construit unidimensionnel.

L'intérêt de ce coefficient est cependant débattu compte tenu de la manière de définir le seuil minimum pour conclure à l'homogénéité de l'échelle (acceptable entre 0,50 et 0,60 pour une étude exploratoire, mais au moins 0,70 voire plus, dans la plupart des autres cas). De plus, la valeur de l'α augmente avec le nombre d'items introduits dans l'échelle, ce qui représente une autre limitation de cette mesure.

Illustrons le calcul de l'α de Cronbach pour l'exemple des sirops de fruit. Le tableau ci-dessous donne les corrélations entre les items fortement corrélés avec le facteur 1.

	Item 1	Item 2	Item 3	Item 10	Item 13
Item 1 : Bon goût de fruit	1,000	0,859	0,803	0,786	0,760
Item 2 : Naturel	**0,859**	1,000	0,777	0,756	0,720
Item 3 : Variété des parfums	**0,803**	**0,777**	1,000	0,706	0,653
Item 10 : Économique (Inverse)	**0,786**	**0,756**	**0,706**	1,000	0,633
Item 13 : Haute qualité	**0,760**	**0,720**	**0,653**	**0,633**	1,000

© Dunod – Toute reproduction non autorisée est un délit.

La moyenne des coefficients de corrélation de tous les couples d'items est $\bar{r}$:

$$\bar{r} = \frac{0,859 + 0,803 + \ldots + 0,633}{10} = 0,745$$

Sachant qu'il y a 5 items, l'alpha de Cronbach pour le facteur 1 est :

$$\alpha = \frac{5 * 0,745}{1 + (5-1) * 0,745} = 0,936$$

Au-delà des normes de seuil évoquées ci-dessus, l'intérêt majeur de l'α de Cronbach est de pouvoir poser un diagnostic sur la qualité de l'échelle construite :

– si sa valeur apparaît trop basse, c'est que les items qui la composent partagent assez peu de variance commune (ou qu'il n'y a pas assez d'items) ;
– si sa valeur est particulièrement élevée, cela traduit probablement une bonne homogénéité de l'échelle. Mais cela peut aussi dissimuler un problème d'ordre méthodologique : si, sur le plan sémantique, les items ne sont pas suffisamment différents (libellés assez voisins par exemple), l'alpha risque d'être élevé sans pour autant couvrir tous les champs possibles de la notion représentée par le facteur. Rappelons de nouveau qu'une analyse factorielle ne peut se substituer aux carences informatives résultant de l'omission de variables fondamentales.

Calculer de nouvelles variables : les scores factoriels

La première fonction de l'analyse factorielle étant de résumer l'information, son application est souvent une phase préliminaire dans un protocole d'étude qui implique d'autres traitements pour remplacer les variables d'origine par des variables composites moins nombreuses : les facteurs. Ils réduisent donc le nombre de critères et focalisent les comparaisons sur les seules informations significatives et communes à l'ensemble des observations. Pour chaque facteur, de nouvelles valeurs calculées remplacent les valeurs des variables initiales, appelées « *score factoriel* ».

Les marques de sirop de fruits seront comparées sur les trois seules dimensions caractéristiques du positionnement à la place des 13 variables initiales. Une échelle des attitudes à l'égard des transports composée de 28 items a été réduite par une ACP à cinq dimensions qui permettent de caractériser les voyageurs de manière synthétique : leurs scores factoriels ont servi de base à la construction du profil de comparaison nécessaire pour effectuer une typologie.

Nombre de logiciels d'analyse factorielle fournissent les scores factoriels, mais pas tous. Il existe différentes méthodes plus ou moins élaborées pour les calculer. Une méthode classique consiste à s'appuyer sur la matrice des poids factoriels, ou *matrice des vecteurs propres* (« eigenvectors ») proposée dans le tableau 4.9.

Le score factoriel est la somme pondérée des valeurs initiales de chaque variable pour un facteur donné. La méthode consiste, pour chaque observation d'un facteur, à multiplier la valeur initiale de la variable par son coefficient factoriel (vecteur propre). La note en facteur pour une observation – ou score factoriel – résulte de la somme de ces valeurs. La formule suivante permet de faire ce calcul :

Tableau 4.9 – Calcul du score factoriel

Procédure de calcul du score factoriel (*SF*) d'une observation
Somme à toutes les variables (k) du produit [Vecteurs Propres (v^k) multiplié par valeurs initiales (x^k)]. Soit : $$SF = \sum_{1}^{K} v^k x^k$$ où $v^k = s_F^k / \sqrt{\lambda_F}$ avec : s_F^k : coefficient de saturation de la variable k dans un facteur donné et $\sqrt{\lambda_F}$: racine de la valeur propre λ_F (lambda) du facteur considéré

Construire des cartes perceptuelles

Les études peuvent suggérer d'autres usages, comme le besoin de construire des cartes factorielles par exemple, dans lesquelles les variables sont projetées dans l'espace des facteurs. Il est également possible de projeter simultanément observations et variables dans un plan factoriel. Mais ce n'est pas la vocation naturelle d'une analyse factorielle. Certains logiciels proposent cependant cette fonction. Pour répondre à cet objectif, il faut transformer les variables en variables fictives (*dummy variable*) afin de calculer leurs coordonnées dans l'espace des facteurs. Seule une représentation par points (observations) et vecteurs (les variables) se justifie. Il faut interpréter avec prudence ce type de projection, trop souvent limitée au plan formé par les deux premiers axes qui ne représentent le plus souvent qu'une petite fraction de la variance commune.

Pour aller plus loin – L'analyse factorielle confirmatoire

L'objectif de l'analyse factorielle confirmatoire est de tester ou confirmer un modèle théorique préétabli. Deux modèles ont été développés à cette fin : Le modèle de causalité (Path analysis) et la modélisation par les équations structurelles sur variables latentes. L'estimation des paramètres des modèles d'équations structurelles peut se faire, soit par l'approche LISREL (*LInear Structural RELationships*) développée par Jöreskog (1970), laquelle consiste à analyser la structure de covariance ; soit par l'approche PLS (*Partial Least Squares regression* ou *Projection to Latent Structure*).

L'ESSENTIEL

Les objectifs d'une ACP

L'analyse factorielle en composantes principales (ACP) a une double fonction :

– réduire l'information initiale pour lui substituer des variables composites et l'utiliser dans d'autres applications statistiques ;

☞

– structurer l'information pour analyser les proximités entre des groupes de variables initiales (les facteurs) et dégager les structures latentes qui rendent compte de ces regroupements ; c'est cette fonction qui est utilisée pour construire des échelles d'attitudes par exemple.

■ *Le déroulement d'une ACP*

Les données sont constituées d'un tableau « observations (en ligne) x variables (en colonne) ». Ces variables, de même statut hiérarchique (pas de distinction entre variables expliquées et explicatives), sont nécessairement métriques et corrélées entre elles.

Les valeurs propres (λ) correspondent aux facteurs : il y a autant de facteurs que de valeurs propres supérieures à 1 quand les variables sont standardisées ou supérieures au ratio 100/p (où p est le nombre total de variables) exprimé en % de variance.

La matrice des saturations (s) ou matrice des corrélations variables-facteurs permet d'identifier les items qui ont généré un facteur et, donc, de lui donner sa signification. Selon la méthode utilisée, l'analyse se fait sur cette matrice, soit avant rotation, soit après rotation : cette solution tend à renforcer la corrélation d'une variable dans un seul facteur, ce qui facilite l'interprétation.

■ *Les précautions liées à l'usage d'une ACP*

Pour chaque variable, la somme du carré de son coefficient de corrélation avec chaque facteur fournit la quantité de sa variance qui est prise en charge par le modèle retenu (100 % si tous les facteurs sont retenus) ; cette somme correspond à la communauté, *i.e.* la quantité de variance partagée. Une variable dont la communauté est inférieure à 50 % (0,50) est mal représentée par le modèle retenu. Si trop de variables ont une faible communauté, cela doit interroger sur la qualité du modèle ou sur la classifiabilité des données brutes.

Le poids d'un facteur correspond à la seule quantité de variance des données dont il rend compte : en aucun cas, il ne peut être associé à une importance de nature subjective ou psychologique. Il faut aussi se méfier des appellations données aux facteurs : elles ne rendent pas toujours bien compte de la réalité qu'ils couvrent ; il faut donc toujours revenir aux informations statistiques, dont les corrélations variables-facteurs.

Section 2 L'ANALYSE FACTORIELLE DES DONNÉES NON-METRIQUES[1]

L'analyse factorielle en composantes principales (ACP) n'opère que sur des données exclusivement métriques. Dans de nombreuses situations cependant, les données nécessaires à l'analyse associent des variables métriques et non métriques. Or, il n'est pas possible de rendre métrique une variable qui ne l'est pas d'origine (à l'exception d'une variable booléenne en 0 et 1, métrique par nature), ni d'utiliser directement l'ACP pour traiter des variables non-métriques.

Plusieurs méthodes existent pour résoudre ce problème statistique lié à la nature des variables : l'Analyse factorielle des Correspondances Multiples (ACM) – objet de cette section – qui s'applique sur des variables nominales et les méthodes d'analyses des similarités et des préférences (« Multi-Dimensionnal Scaling » ou MDS) qui s'appliquent sur des données ordinales (il n'en est pas traité dans cet ouvrage). Ces deux classes de méthodes sont souvent utilisées, en marketing par exemple pour étudier le positionnement de plusieurs marques ou produits comparés sur des critères non métriques.

Dans ses principes mathématiques et méthodologiques, l'ACM est comparable à l'AFC. Mais son usage implique que toutes les variables soient nominales : il faut donc transformer toutes les variables métriques présentes pour construire ensuite le tableau disjonctif complet dans lequel chaque modalité des variables d'origine devient une variable spécifique, indiquant si une observation possède (1) ou non (0) le caractère porté par chaque modalité. Ce tableau permet de construire le tableau de contingence qui contient l'ensemble des variables-modalités (appelé « tableau de Burt ») sur lequel porte le calcul statistique (en fait, la réalisation d'une ACP).

Cas d'illustration

Téléphones sans fils

Un catalogue propose un certain nombre de modèles de téléphones sans fils, décrits sur un certain nombre de critères concernant leurs caractéristiques techniques (l'autonomie en veille, la portée, le poids…) et leur niveau d'équipement (les fonctions « mains-libres » et « baby-sitting », le rétroéclairage du clavier et un certain nombre de services divers), ainsi que leur prix de vente. Cette liste de critères est complétée par deux notes : l'une (sur 10), technique, évalue la portée réelle du combiné et l'autre (sur 20), plus globale et subjective, évalue l'agrément à l'usage (performance des équipements, confort d'utilisation et esthétique). Le tableau 4.10 récapitule ces informations pour quatorze appareils présélectionnés. ☞

1. La lecture de cette section implique d'avoir pris connaissance de la partie consacrée à l'analyse des correspondances simples (ou AFC) présentée dans le chapitre précédent.

© Dunod – Toute reproduction non autorisée est un délit.

☞ Or, on constate de gros écarts de prix (presque dans un rapport de 1 à 4). Aussi, avant de passer commande, l'acheteur potentiel se demande si ces écarts ne seraient pas plus liés à la marque qu'au niveau d'équipement : la notoriété et la réputation de certaines d'entre elles pouvant influer sur le prix. Si c'est avéré, alors pourquoi payer plus cher un combiné qui apporte autant de commodités qu'un appareil à meilleur marché ?

1 Construire le tableau des données adapté à l'ACM

1.1 La transformation des variables

Le tableau 4.10 fournit les données brutes. Trois variables décrivent la possession (1) ou non (0) d'une caractéristique : les fonctions « mains-libres », « baby-sitting », et « rétroéclairage » ; composées de deux modalités, elles sont non-métriques et peuvent être conservées en l'état. Les six autres variables, qui sont métriques, doivent être transformées pour être exploitables avec une ACM.

Tableau 4.10 – Tableau des données brutes (cas « téléphones sans fils »)

	Prix (€)	Poids (gr)	Autonomie de veille (en h.)	Éclairage du clavier	Fonction baby-sitting	Fonction mains-libres	Nombre de services	Portée réelle (note / 10)	Agrément (note / 20)
VERSA 700	80	92	100	0	0	1	1	8	15,0
MATRA 530	70	131	120	0	0	1	3	8	16,0
LEXI DP232	80	113	80	0	0	0	2	7	14,5
LOGY LP532	60	119	100	0	0	1	2	6	14,5
PANA 432	70	115	120	0	0	0	2	7	15,5
PHIL 511	130	99	200	0	1	1	3	8	16,0
SAG D20T	90	139	160	0	0	1	1	7	15,0
GIGA S430	50	117	170	0	0	0	1	9	16,0
VERSA 1600	90	92	100	1	0	1	3	8	16,0
MATRA 850	90	111	150	0	0	1	2	8	15,0
LEXI DP611	120	124	70	1	1	1	2	8	15,0
PHIL 617	200	108	240	1	1	0	4	8	16,5
SAG D85C	220	140	180	1	1	1	4	8	16,0
GIGA S740	200	109	300	1	1	0	4	8	15,5

Chaque variable métrique doit être décomposée en un certain nombre de classes qui deviennent les modalités de la variable transformée. L'analyse de la distribution de leurs valeurs donne une idée du nombre de classes à fixer (*cf.* chapitre 1, section 2, § 1 « Construire le tableau de distribution ») ; et ce nombre doit respecter

la dispersion des valeurs. Cependant, pour que l'analyse reste pertinente, il faut qu'il y ait un minimum d'observations par classe, afin que les modalités retenues permettent encore de discriminer les objets décrits. Pour partie, l'intérêt de l'analyse et la facilité d'interprétation dépendent de la pertinence des variables utilisées, mais aussi de la qualité de la transformation des variables métriques initiales.

Par ailleurs, sachant que :

– la contribution de la modalité *m* d'une variable à l'inertie totale est d'autant plus forte que sa fréquence est faible,
– la contribution d'une variable d'origine *p* est d'autant plus forte que le nombre de ses modalités *m* est élevé,

il est donc recommandé d'utiliser des variables ayant un nombre comparable de modalités avec une répartition des observations équilibrée entre ces modalités.

Compte tenu des contraintes évoquées, mais aussi des données disponibles, quatre variables (Poids/Autonomie/Portée/Agrément) ont été décomposées en deux classes, Prix en trois classes et Nombre de services en quatre classes (sans doute celle qui pèse le plus). Mais cette décomposition fournit pour chaque modalité *m* des différentes variables *p* un nombre d'observations équilibré.

Dans le tableau 4.11 qui présente les données métriques transformées, les quatorze modèles sont présentés en ligne ($n = 14$) et chaque colonne contient leurs valeurs (variant de 1 à 4 selon le nombre de *m* modalités) sur les caractères introduits dans l'analyse ($p = 9$). Par l'extrait qu'il contient, le tableau 4.12 montre ce qu'est un tableau disjonctif complet.

Tableau 4.11 – Tableau des variables transformées en classes et rendues non métriques (cas « téléphones sans fils »)

	Prix	Poids	Autonomie	Éclairage	Babysit	Mains libres	Nb services	Portée	Agrément
VERSA 700	1	1	1	1	1	2	1	2	1
MATRA 530	1	2	1	1	1	2	3	2	2
LEXI DP232	1	1	1	1	1	1	2	1	1
LOGY LP532	1	2	1	1	1	2	2	1	1
PANA 432	1	2	1	1	1	1	2	1	1
PHIL 511	3	1	2	1	2	2	3	2	2
SAG D20T	2	2	2	1	1	2	1	1	1
GIGA S430	1	2	2	1	1	1	1	2	2
VERSA 1600	2	1	1	2	1	2	3	2	2
MATRA 850	2	1	2	1	1	2	2	2	1
LEXI DP611	2	1	1	2	2	2	2	2	1
PHIL 617	3	1	2	2	2	1	4	2	2
SAG D85C	3	2	2	2	2	2	4	2	2
GIGA S740	3	2	2	2	2	1	4	2	1

© Dunod – Toute reproduction non autorisée est un délit.

Tableau 4.12 – Tableau de Burt

	Poids_1	Poids_2	Autonomie_1	Autonomie_2	Éclairage_1	Éclairage_2	Babysit_1	Babysit_2	Mainslibr_1	Mainslibr_2	Nb Servic_1	Nb Servic_2	Nb Servic_3	Nb Servic_4	Portée_1	Portée_2	Agrément_1	Agrément_2
Poids_1	7	0	4	3	4	3	4	3	2	5	1	3	2	1	1	6	4	3
Poids_2	0	7	3	4	5	2	5	2	3	4	2	2	1	2	3	4	4	3
Autonomie_1	4	3	7	0	5	2	6	1	2	5	1	4	2	0	3	4	5	2
Autonomie_2	3	4	0	7	4	3	3	4	3	4	2	1	1	3	1	6	3	4
Éclairage_1	4	5	5	4	9	0	8	1	3	6	3	4	2	0	4	5	6	3
Éclairage_2	3	2	2	3	0	5	1	4	2	3	0	1	1	3	0	5	2	3
Babysit_1	4	5	6	3	8	1	9	0	3	6	3	4	2	0	4	5	6	3
Babysit_2	3	2	1	4	1	4	0	5	2	3	0	1	1	3	0	5	2	3
Mainslibr_1	2	3	2	3	3	2	3	2	5	0	1	2	0	2	2	3	3	2
Mainslibr_2	5	4	5	4	6	3	6	3	0	9	2	3	3	1	2	7	5	4
Nb servic_1	1	2	1	2	3	0	3	0	1	2	3	0	0	0	1	2	2	1
Nb servic_2	3	2	4	1	4	1	4	1	2	3	0	5	0	0	3	2	5	0
Nb servic_3	2	1	2	1	2	1	2	1	0	3	0	0	3	0	0	3	0	3
Nb servic_4	1	2	0	3	0	3	0	3	2	1	0	0	0	3	0	3	1	2
Portée_1	1	3	3	1	4	0	4	0	2	2	1	3	0	0	4	0	4	0
Portée_2	6	4	4	6	5	5	5	5	3	7	2	2	3	3	0	10	4	6
Agrément_1	4	4	5	3	6	2	6	2	3	5	2	5	0	1	4	4	8	0
Agrément_2	3	3	2	4	3	3	3	3	2	4	1	0	3	2	0	6	0	6

1.2 Variables « actives » et « illustratives » : une différence de statut

Dans le cas proposé, l'étude a pour objectif de vérifier avec une ACM si le prix de vente d'un combiné est ou non lié à la marque : pour un même niveau d'équipement et de services, le prix serait d'autant plus élevé que la marque est bien connue et son image valorisée. Si c'est avéré, il faut s'attendre à ce que les combinés commercialisés par les marques les plus prisées soient associés aux niveaux de prix les plus élevés et, inversement, les moins connues aux prix les plus bas.

Pour vérifier ce type d'hypothèse, il faut en quelque sorte neutraliser la variable « Prix » pour qu'elle n'intervienne pas dans le calcul de l'inertie dont les facteurs rendent compte, inertie qui ne doit alors dépendre que des seuls attributs techniques.

L'analyse factorielle des correspondances multiples (tout comme l'analyse des correspondances simples d'ailleurs) permet de différencier les variables selon le rôle qu'elles jouent dans l'étude conduite. Les variables (ou les modalités dans un

tableau croisé) qui interviennent directement dans la genèse des facteurs sont des variables (ou des modalités) « actives » ; celles qui n'interviennent pas dans le calcul des inerties et qui ne contribuent donc pas à l'émergence des axes sont des « variables muettes » (ou « *dummy variables* »), dites « illustratives » : c'est le cas du « Prix » dans l'exemple traité. Cependant, le calcul de leurs coordonnées (voir de leurs valeurs tests) permet de les situer sur chaque axe : elles constituent donc une aide à l'interprétation des facteurs.

> Par exemple, dans de nombreuses études portant sur la consommation, les variables socio-démographiques sont utilisées comme variables illustratives pour mieux cerner les segments de clientèle, voire émettre des hypothèses de ciblage qui pourront être ensuite éprouvées par une analyse typologique.

Pour aller plus loin – Quelques compléments spécifiques à l'ACM

Ce qu'est un tableau disjonctif complet et comment il est construit

Le tableau disjonctif complet (TDC) est obtenu en recodant chacune des modalités des p variables qualitatives initiales en autant de variables binaires *(0/1)* en créant des variables qualitatives avec le code 1 pour la modalité observée et *0* pour toute autre modalité. Ainsi, le nombre des variables binaires (colonnes) d'un tableau disjonctif complet m est égal à la somme des modalités de toutes les variables initiales. Le TDC dispose donc de p lignes (individus) et de m colonnes (modalités binaires). Le tableau ci-dessous correspond au tableau disjonctif complet obtenu à partir du tableau des huit variables actives du cas « téléphones sans fils » transformées en classes (*cf.* tableau 4.11) et rendues non métriques.

	Poids -	Poids +	Autonomie -	Autonomie +	Éclairage non	Éclairage oui	Baby-sitter non	Baby-sitter oui	Mains libres non	Mains libres oui	Nb services -	Nb services =	Nb services +	Nb services ++	Portée -	Portée +	Agrément -	Agrément +	Total
VERSA 700	1	0	1	0	1	0	1	0	0	1	1	0	0	0	0	1	1	0	***8***
MATRA 530	0	1	1	0	1	0	1	0	0	1	0	0	1	0	0	1	0	1	***8***
LEXI DP232	1	0	1	0	1	0	1	0	1	0	0	1	0	0	1	0	1	0	***8***
LOGY LP532	0	1	1	0	1	0	1	0	0	1	0	1	0	0	1	0	1	0	***8***
PANA 432	0	1	1	0	1	0	1	0	1	0	0	1	0	0	1	0	1	0	***8***
PHIL 511	1	0	0	1	1	0	0	1	0	1	0	0	1	0	0	1	0	1	***8***
SAG D20T	0	1	0	1	1	0	1	0	0	1	1	0	0	0	1	0	1	0	***8***
GIGA S430	0	1	0	1	1	0	1	0	1	0	1	0	0	0	0	1	0	1	***8***
VERSA 1600	1	0	1	0	0	1	1	0	0	1	0	0	1	0	0	1	0	1	***8***
MATRA 850	1	0	0	1	1	0	1	0	0	1	0	1	0	0	0	1	1	0	***8***
LEXI DP611	1	0	1	0	0	1	0	1	0	1	0	1	0	0	0	1	1	0	***8***
PHIL 617	1	0	0	1	0	1	0	1	1	0	0	0	0	1	0	1	0	1	***8***
SAG D85C	0	1	0	1	0	1	0	1	0	1	0	0	0	1	0	1	0	1	***8***
GIGA S740	0	1	0	1	0	1	0	1	1	0	0	0	0	1	0	1	1	0	***8***
Total	***7***	***7***	***7***	***7***	***9***	***5***	***9***	***5***	***5***	***9***	***3***	***5***	***3***	***3***	***4***	***10***	***8***	***6***	

☞

© Dunod – Toute reproduction non autorisée est un délit.

La somme marginale des lignes du TDC est toujours égale au nombre de variables initiales p (8 dans le cas des téléphones). La somme marginale des variables pouvant aller de *1* à n indique le nombre d'observations possédant une modalité (propriété) donnée. À titre d'exemple, la plupart des téléphone (10 sur 14) possèdent la propriété « Portée + » alors que très peu de téléphones (3 sur 14) possèdent la propriété « Nb services ++». C'est donc cette propriété qui distingue des autres. Cette distinction - la probabilité pour qu'une modalité soit possédée par les observations - notée p_k pour une modalité donnée (propriété) j est mesurée en divisant sa valeur marginale n_j du TDC par le nombre d'observations : $p_j = n_j/n$

Dans l'exemple des téléphones, tous les 1 de la variable « Nb services ++» deviennent 1/(7/14)=2 alors que les 1 propriétés « Portée + » deviennent beaucoup plus faibles 1/(10/14)=1,4. De cette façon les valeurs du profil moyen sont toutes égales à 1. Mais l'ACM, travaillant avec des valeurs centrées, il faut alors retrancher 1 aux valeurs transformées comme fait ci-dessus. Ainsi, tous les 1 de la variable « Nb services ++» deviennent 1/(7/14)-1=1 alors que les 1 de « Portée + » deviennent beaucoup plus faibles 1/(10/14)-1=0,4. L'ACM effectue alors une analyse factorielle en composantes principale appliquée sur ces données centrées et rendues métriques. Tout comme l'ACP, l'ACM étudie séparément l'espace « individus-observations » et l'espace variables selon la procédure déjà expliquée pour l'ACP.

En ACM, l'inertie totale – qui vaut la somme de l'inertie de chaque point du nuage - est obtenue en divisant le nombre de colonnes du TDC par le nombre des variables initiales p en retranchant 1 :

$$Inertie\ totale = \frac{m}{p} - 1$$

Ce rapport m/p représente le nombre moyen de modalités par variable initiale. En ACM, l'inertie totale est donc égale au nombre moyen de modalités moins une (alors qu'en AFC, elle est égale au Khi-deux du tableau de contingence divisé par le nombre d'individus statistiques).

À la différence de l'ACP, en ACM on ne retient que les composantes pour lesquelles les valeurs propres sont supérieures à $1/p$ *i.e.* à un divisé par le nombre des variables initiales. Il y a d'autres approches basées sur des tests statistiques qui permettent de sélectionner les composantes principales significatives. Une de ces approches fait appel aux corrélations qui existent entre les variables initiales et les composantes principales : une autre, mise en place par Lebart, Morineau et Warwick (1984)[1] est intégrée par le logiciel SPAD. Le développement de ces approches sort du cadre de cet ouvrage.

L'ACM peut aussi être considérée comme une AFC appliquée sur un tableau de Burt. Le tableau de Burt est une matrice carrée symétrique construite en croisant les variables prises deux par deux (*cf.* tableau 4.12). Il peut être facilement obtenu à partir du tableau disjonctif complet. Par exemple, parmi les téléphones sans fils étudiés (voir tableau ci-après), sept ont un poids faible « Poids - » et les sept autres ont un poids important « Poids + ». En construisant le tableau de contingence qui croise la variable « Poids » avec elle-même, on obtient :

		Poids		Total
		Poids -	Poids +	
Poids	Poids -	7	0	7
	Poids +	0	7	7
Total		7	7	14

De même, le tableau de contingence entre les deux variables « Poids » et « NbServices » donne :

		Poids		Total
		Poids -	Poids +	
NbServices	Nb servic -	1	2	3
	Nb servic =	3	2	5
	Nb servic +	2	1	3
	Nb servic ++	1	2	3
Total		7	7	14

Statistiquement parlant, le tableau de Burt est le produit matriciel entre la matrice du disjonctif complet et sa transposée. Les tableaux de contingence croisant les variables avec elles-mêmes se trouvent sur la diagonale du tableau de Burt (*cf.* tableau 4.12). La somme des diagonales des tableaux de contingence (trace) est toujours égale au nombre d'unités statistiques étudiées (14 dans le cas « Téléphones »). L'ACM utilise la matrice diagonale qui ne reprend que les sous-matrices qui se trouvent sur la diagonale du tableau de Burt en mettant 0 partout ailleurs. Le tableau de Burt est moins volumineux que le tableau disjonctif complet parce qu'il contient moins d'information. L'inconvénient est qu'il ne permet pas de tracer le nuage de points des individus statistiques.

2 Les différentes étapes pour analyser les résultats d'une ACM

Les bases mathématiques de l'analyse factorielle des correspondances multiples sont identiques à celles de l'analyse des correspondances simples (*cf.* chapitre 3, section 2, § 1.2), les calculs fournissent donc le même type de résultats :

– les valeurs propres correspondant à une part de la variance totale prise en charge par le modèle : elles permettent de déterminer le nombre d'axes significatifs et la qualité explicative du modèle retenu ;
– les différentes matrices contenant les informations nécessaires à l'interprétation des axes, à savoir :
  – les *contributions absolues :* poids d'une modalité dans l'émergence d'un axe ;

1. Lebart, L., Morineau, A. et Warwick, K.W. (1984). *Multivariate Descriptive Statistical Analysis, Correspondence Analysis and Related Techniques for Large Matrices*, Wiley, New-York.

© Dunod – Toute reproduction non autorisée est un délit.

  - les *contributions relatives* qui indiquent la part de variance d'une modalité prise en charge par chaque facteur ;
  - les *coordonnées des modalités* sur les axes qui permettent d'identifier et d'évaluer les rapports de proximité ou d'opposition entre ces modalités ;
- les graphiques projetant les modalités des variables et des observations sur les axes retenus pour visualiser les résultats.

Pour analyser et interpréter les résultats d'une ACM, la démarche reste identique à celle développée pour l'AFC (voir chapitre 3, section 2, § 2). Il faut cependant noter que l'étude des proximités entre les modalités en ligne (« observations ») et les modalités en colonne (« attributs ») n'est possible et correcte qu'avec une projection de type barycentrique, où les attributs sont projetés dans l'espace des observations (individus ou objets selon le contenu des variables étudiées). Une projection standard classique ne permet plus d'analyser directement ces proximités.

2.1 Déterminer les valeurs seuils pour identifier les axes significatifs

Pour qu'une valeur propre (et donc l'axe qui lui correspond) soit significative, il faut que sa valeur soit supérieure au ratio « 100/*p* », lequel correspond à la moyenne des valeurs propres. Dans ce ratio, *p* est le nombre de variables d'origine « actives » dans le calcul de l'inertie totale ; 100 est le pourcentage d'inertie totale (inertie ou variance totale = 100 %).

On peut aussi s'appuyer sur l'histogramme des valeurs propres dont chaque inflexion de la courbe correspond à un seuil de significativité. Certains logiciels (tel SPAD) fournissent un indicateur pour optimiser le nombre d'axes significatifs. Rappelons que pour être opérationnel, le modèle factoriel retenu doit rendre compte avec le minimum de facteurs du maximum d'inertie (si possible supérieur à 60 %).

Par ailleurs, il est démontré que le nombre total de valeurs propres vaut « *m-p* », où *m* est le nombre de modalités-variables « actives » composant le tableau disjonctif. On peut en déduire la valeur totale de l'inertie qui correspond à la somme des valeurs propres : (*m/p*)*-1*. Ces éléments permettent de construire le modèle.

> Pour le cas « téléphones sans fils », il y a 9 variables d'origine (*p*), mais la variable Prix est « illustrative » : il n'y a donc que 8 variables « actives » pour le calcul de l'inertie totale ; par ailleurs, ces 8 variables ont été restructurées en 18 modalités actives ; l'inertie totale vaut donc : (*m/p*) *-1*, soit (18/8)-1 = (2,25-1) = 1,25.
>
> La valeur moyenne des valeurs propres qui doit être dépassée (exprimée en %) vaut (100/*p*). Sachant que dans le cas traité, il y a 8 variables initiales « actives », soit *p* = 8, ce ratio vaut donc 100/8, soit 12,5 %. Les trois valeurs propres dépassant cette valeur moyenne sont donc significatives et définissent les trois facteurs à retenir.

Ces valeurs sont directement lisibles sur la figure 4.2 : l'examen de l'histogramme permet de constater une forte rupture entre la troisième et la quatrième valeur propre et confirme la solution avec 3 facteurs, particulièrement économique : elle rend compte de plus des 2/3 de l'inertie totale, soit 68,71 %.

2.2 Déterminer les valeurs seuils pour identifier les modalités significatives

Une modalité colonne (« attributs ») est significative dès lors que la valeur-test calculée évaluant la force de la relation modalité-facteur est égale ou supérieure à une certaine valeur. Selon le logiciel utilisé, cette valeur peut être différente car fondée sur deux approches différentes.

Une méthode est fondée sur la corrélation entre une modalité et un axe : la significativité du coefficient η^2 (qui est un coefficient de détermination – *cf.* chapitre 2) est testée avec l'analyse de variance : la valeur seuil du test F est 4 ; toute valeur test égale ou supérieure à ce seuil indique une relation significative entre l'axe et la variable-modalité.

Une autre méthode est utilisée par le logiciel SPAD : elle repose sur le principe d'un test d'hypothèse nulle classique ; dès lors que sa valeur absolue vaut 2 (en fait supérieure à 1,96), le lien entre l'axe et la variable-modalité est significatif. Le cas d'illustration utilisé ayant été traité avec SPAD, c'est cette valeur qui est retenue ici pour interpréter les résultats (tableau 4.16).

Dans l'exemple traité, toutes les modalités sont liées au moins à un facteur. Par exemple, seules les variables « poids » et « mains-libres » ne sont pas liées avec l'axe 1, mais la première l'est avec l'axe 3 et la seconde avec l'axe 2. L'axe 1 est lié significativement avec six des huit variables ; l'axe 3 l'est avec deux d'entre elles : la variable poids et les deux modalités les plus restreintes de la variable « services ».

À défaut de valeurs-tests pour les lignes (« observations »), on procède selon le même principe qu'en AFC : l'ensemble des observations représentant 100 % de la variance des lignes, la valeur moyenne de la variance qui doit être dépassée vaut donc $100/n$ où n est le nombre d'observations (soit 100/14 = 7,14 dans le cas).

2.3 Interpréter les axes

La méthode consiste à considérer, axe par axe, les modalités dont les contributions absolues sont significatives (sur la base des valeurs-tests tel que précisé ci-dessus) et, après examen du signe de leurs coordonnées, décider si ces modalités sont liées positivement ou négativement avec l'axe. L'analyse des contributions relatives permet de vérifier si chacun des axes représente bien la variance de chacune des modalités-variables. Cette double analyse permet donc d'interpréter les facteurs et de valider la qualité « explicative » du modèle final.

© Dunod – Toute reproduction non autorisée est un délit.

```
APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION ..   1.2500
                                     SOMME DES VALEURS PROPRES ....   1.2500
```

- TABLEAU DES VALEURS PROPRES avec l'histogramme de leur distribution et les seuils de concavité (critère de Kaiser)

NUMERO	VALEUR PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	**0.4184**	**33.47**	**33.47**	*****

Tableau 4.13 – Statistiques de l'ACM sur les attributs pour les trois premiers axes

Libellé des modalités actives	Effectif	Distance à l'origine	Contributions (CTR)			Cosinus carrés (COS^2)			Coordonnées (COOR)		
			Axe 1	Axe 2	Axe 3	Axe 1	Axe 2	Axe 3	Axe 1	Axe 2	Axe 3
POIDS											
C5=1 -	7	1,0000	0,23	3,86	10,21	0,02	0,15	0,32	-0,12	0,39	-0,57
C5=2 +	7	1,0000	0,23	3,86	10,21	0,02	0,15	0,32	0,12	-0,39	0,57
VEILLE											
C6=1 -	7	1,0000	4,68	2,57	7,60	0,31	0,10	0,24	0,56	0,32	-0,49
C6=2 +	7	1,0000	4,68	2,57	7,60	0,31	0,10	0,24	0,00	0,00	0,00
ÉCLAIRAGE											
C7=0 -	9	0,5556	6,18	0,59	3,57	0,58	0,03	0,16	0,57	0,13	0,30
C7=1 +	5	1,8000	11,12	1,06	6,43	0,58	0,03	0,16	-1,02	-0,24	-0,53
BABY-SITTING											
C8=0 -	9	0,5556	7,10	1,03	1,80	0,67	0,06	0,08	0,61	0,18	0,21
C8=1 +	5	1,8000	12,78	1,85	3,24	0,67	0,06	0,08	-1,09	-0,32	-0,38
MAINS LIBRES											
C9=0 -	5	1,8000	0,05	14,64	0,19	0,00	0,44	0,00	-0,07	-0,89	0,09
C9=1 +	9	0,5556	0,03	8,13	0,10	0,00	0,44	0,00	0,04	0,50	-0,05
SERVICES											
C10=1 -	3	3,6667	1,77	0,00	25,45	0,08	0,00	0,51	0,53	-0,02	1,37
C10=2 =	5	1,8000	7,02	2,01	16,25	0,37	0,06	0,40	0,81	-0,33	-0,85
C10=3 +	3	3,6667	1,06	28,14	0,05	0,05	0,70	0,00	-0,41	1,60	0,06
C10=4 ++	3	3,6667	13,85	11,60	0,00	0,59	0,29	0,00	-1,47	-1,03	-0,02
PORTÉE											
C11=1 -	4	2,5000	12,12	5,32	0,02	0,57	0,15	0,00	1,19	-0,60	-0,04
C11=2 +	10	0,4000	4,85	2,13	0,01	0,57	0,15	0,00	-0,48	0,24	0,01
AGRÉMENT											
C12=1 -	8	0,7500	5,26	4,55	3,11	0,41	0,21	0,11	0,56	-0,39	-0,29
C12=2 +	8	1,3333	7,01	6,07	4,15	0,41	0,21	0,11	-0,74	0,53	0,39

Modalités illustratives			Coordonnées		
PRIX					
C4=1 -	6	1,3333	0,77	0,01	0,23
C4=2 =	4	2,5000	0,16	0,40	-0,37
C4=3 +	4	2,5000	-1,31	-0,42	0,03

Illustration avec le cas « téléphones » (*cf.* colonnes « contributions » et « coordonnées » des tableaux 4.13 et 4.14) :

L'**axe 1** (33,47 % de l'inertie totale) associe une variable subjective (évaluation globale ou *agrément*) aux potentialités des combinés : l'agrément est d'autant plus marqué que le niveau d'équipement des modèles est élevé (« | » signifie symétriquement opposés sur les pôles d'un axe) : agrément : C12 = 2 | 1 avec C10 = 4 | 2 : nombre de services ; C6 = 3 | 1 et C11 = 2 | 1 : durée de veille et portée ; C7 = 1 | 0 : éclairage du clavier ; C8 = 1 | 0 : fonction baby-sitting).

Par ailleurs, l'**axe 1** oppose les trois modèles PHIL617 / SAGD85C / GIGAS740 à trois autres modèles LOGYPI532 / LEXIDP232 / PANA432. La superposition des proximités entre les lignes d'une part et entre les colonnes d'autre part, montre que cet axe oppose les modèles les mieux équipés et les plus appréciés à ceux qui le sont le moins (*cf.* figure 4.3).

L'**axe 2** (19,49 % de l'inertie totale) est généré par l'opposition entre les deux modalités de la fonction « mains libres » (C9 = 0 | 1) et le nombre élevé de services proposés (C10 = 4 | 3). Cet axe distingue un appareil (GIGA S740) de trois autres modèles (PHIL 511 / VERSA 1600 / MATRA 530).

Enfin, l'**axe 3** (15,75 % de l'inertie totale) associe les variables « poids » (C5 = 1| 2) et « nombre de services » faible (C10 = 1 | 2). Il associe deux appareils (GIGAS430 / SAGD20T) opposés à un troisième (LEXIDP611).

Tableau 4.14 – Statistiques de l'ACM sur les « individus » pour les premiers trois axes

Analyse des individus actifs				Contributions			Cosinus carrés			Coordonnées	
Modèles	Poids relatif	Distance origine	Axe 1	Axe 2	Axe 3	Axe 1	Axe 2	Axe 3	Axe 1	Axe 2	Axe 3
VERSA700	7,14	1,06	3,24	3,37	0,69	0,18	0,11	0,02	0,44	0,34	0,14
MATRA530	7,14	1,13	0,05	18,09	2,88	0,00	0,54	0,07	0,05	0,79	0,28
LEXI DP232	7,14	1,25	10,72	2,74	7,70	0,50	0,07	0,17	0,79	-0,31	-0,46
LOGI PY532	7,14	1,09	12,65	0,66	1,19	0,68	0,02	0,03	0,86	-0,15	-0,18
PANA 432	7,14	1,25	12,05	7,39	0,72	0,57	0,20	0,02	0,84	-0,50	-0,14
PHIL 511	7,14	1,29	4,99	14,22	0,19	0,23	0,38	0,00	-0,54	0,70	0,07
SAG D20T	7,14	1,32	5,93	1,57	18,72	0,26	0,04	0,39	0,59	-0,23	0,72
GIGA S430	7,14	1,29	0,00	0,55	33,79	0,00	0,01	0,72	0,00	-0,14	0,97
VERSA 1600	7,14	1,29	1,56	23,08	2,67	0,07	0,61	0,06	-0,30	0,89	-0,27
MATRA 850	7,14	0,83	1,28	0,29	1,61	0,09	0,01	0,05	0,27	0,10	-0,21
LEXI DP611	7,14	1,14	0,36	0,05	28,42	0,02	0,00	0,69	-0,15	0,04	-0,89
PHIL 617	7,14	1,60	19,67	5,07	0,75	0,72	0,11	0,01	-1,07	-0,42	-0,14
SAG D85C	7,14	1,44	17,25	1,99	0,67	0,70	0,05	0,01	-1,01	-0,26	0,14
GIGAS 740	7,14	1,53	10,26	20,94	0,01	0,39	0,47	0,00	-0,78	-0,85	-0,02

Les cellules en grisé correspondent aux modèles dont le ratio 100 / 14 > 7,14

Évaluation de la validité du modèle : sa capacité à rendre compte des données

Le tableau 4.15 rend compte de la variance prise en charge par les trois facteurs du modèle retenu pour chacune des modalités actives (rappelons que la variance totale d'une modalité vaut 1 ou 100 %). Si toutes les modalités des variables actives sont prises en charge par le modèle en trois facteurs, on note que deux modèles (VERSA 700 et MATRA 850) ne le sont pas. En fait, le quatrième facteur les prend en charge, en liaison avec le seul attribut « poids », attribut qui a aussi généré l'axe 3. L'axe 4 et les suivants ne sont qu'un affinement de l'analyse liée aux cinq modèles dont les deux premiers axes ne rendent pas compte (VERSA 700, MATRA 850, LEXI DP611, GIGA S430, SAG D20T), modèles par ailleurs proches du centre de gravité, correspondant donc à la moyenne des produits.

Tableau 4.15 – Variance totale des modalités expliquée par le modèle en 3 facteurs (cosinus-carré cumulés)

COS^2 Cumulé sur les 3 axes							
Modalités «variables»				Modalités «observations»			
poids_1	0,4875	nb servic_1	0,5860	VERSA700	0,3051	GIGA S430	0,7374
poids_2	0,4875	nb servic_2	0,8246	MATRA530	0,6168	VERSA 1600	0,7385
autonomie_1	0,6529	nb servic_3	0,7443	LEXI DP232	0,7498	MATRA 850	0,1566
autonomie_2	0,6529	nb servic_4	0,8777	LOGI PY532	0,7305	LEXI DP611	0,7084
écairage_1	0,7685	portée_1	0,7135	PANA 432	0,7855	PHIL 617	0,8409
écairage_2	0,7685	portée_2	0,7135	PHIL 511	0,6070	SAG D85C	0,7590
babysit_1	0,8008	Agrément_1	0,7320	SAG D20T	0,6931	GIGAS 740	0,8615
babysit_2	0,8008	Agrément_2	0,7320	Les valeurs en grisé correspondent aux modalités dont le cosinus carré est inférieur à 50 % : le modèle en 3 facteurs ne rend pas bien compte de ces modalités.			
mainslibr_1	0,4511						
mainslibr_2	0,4511						

Ce type d'analyse a pour objectif de faire ressortir ce qui se distingue le plus de la moyenne. Plus un objet ou une modalité sont excentrés par rapport à l'origine des axes, plus leur caractère est original et plus leur profil se distingue de la masse. Cette particularité présente un risque : si une observation est très originale (voire aberrante) ou si une modalité-variable se distingue par un effectif très éloigné de la moyenne, cela peut générer un axe spécifique et perturber la cohérence de l'ensemble en rapprochant les autres points du centre de gravité, les rendant moins distinguables. Il importe donc de vérifier que les données traitées ne comportent pas d'observations atypiques (ce qui, dans l'exemple, pourrait être le cas du modèle MATRA 850). Si c'est le cas, il est prudent d'effectuer une nouvelle ACM en traitant les observations ou les modalités très particulières en données « illustratives » pour qu'elles ne polluent, ni le calcul de l'inertie et des valeurs propres, ni l'interprétation des axes.

© Dunod – Toute reproduction non autorisée est un délit.

Tableau 4.16 – Valeurs tests des modalités « attribut » pour les trois premiers axes

Variables :		AXE 1	AXE 2	AXE 3	Variables :		AXE 1	AXE 2	AXE 3
Poids	C5 = 1	- 0,4	1,4	- 2,0	Mains-libres	C 9 = 0	- 0,2	- 2,4	0,2
	C5 = 2	0,4	- 1,4	2,0		C 9 = 1	0,2	2,4	-0,2
Veille	C6 = 1	2,0	1,1	- 1,8	Portée	C11= 1	2,7	- 1,4	- 0,1
	C6 = 2	- 2,0	- 1,1	1,8		C11= 2	- 2,7	1,4	0,1
Éclairage	C7 = 0	2,7	0,6	1,4	Agrément	C12= 1	2,3	- 1,6	- 1,2
	C7 = 1	- 2 ,7	- 0,6	- 1,4		C12= 2	- 2,3	1,6	1,2
Baby-sitting	C8 = 0	2,9	0,9	1,0	Service	C10= 1	1,0	0,0	2,6
	C8 = 1	- 2,9	- 0,9	- 1,0		C10= 2	2,2	- 0,9	- 2,3
Prix	C4 = 1	2,4	0,0	0,7		C10= 3	- 0,8	3,0	0,1
	C4 = 2	0,4	0,9	- 0,8		C10= 4	- 2,8	- 1,9	0,0
	C4 = 3	- 3.0	- 1.0	0.1	« Prix » est une variable illustrative et non active				

Les cellules en grisé soulignent pour chaque axe, les valeurs test au moins égales à 2.

Vérification de l'hypothèse relative à l'effet marque sur le prix

La variable **Prix** n'est significativement liée qu'au premier facteur par ses valeurs extrêmes (C4.1 et C4.3, *cf.* valeurs tests calculées, tableau 4.16). Les modèles les plus chers sont aussi les mieux équipés et les plus appréciés, les moins chers étant les moins équipés. D'un point de vue économique, ces constats sont logiques : on peut donc conclure qu'il n'y a pas un effet marque important. Sur cet échantillon de combinés testés, il semble bien qu'il y a un lien marqué entre niveau de gamme et prix. Par ailleurs, la dispersion de ces appareils dans les plans composés par les trois axes retenus donne à penser qu'il y a trois ou quatre groupes de modèles dont la pertinence pourrait être vérifiée en effectuant une typologie ultérieurement (la réalisation de cette typologie sur la base des trois facteurs confirme bien l'existence de quatre groupes dont l'un d'eux associe fortement les modèles les mieux équipés aux prix les plus élevés).

2.4 Les limites d'une interprétation directe fondée sur les cartes perceptuelles

La synthèse qui précède s'appuie sur l'analyse des tableaux de résultats fournis par l'analyse des correspondances multiples, en ne retenant que les seules modalités significatives dans chaque axe retenu (*cf.* tableaux 4.13 et 4.14) et contenues à l'intérieur des zones dessinées dans les plans reproduits. Cet exemple illustre très bien les risques encourus par une analyse qui ne serait fondée que sur la lecture des graphiques, sans tenir compte de la significativité des résultats obtenus (tableau des contributions absolues) et de la qualité de projection des modalités et des observations (tableau des contributions relatives).

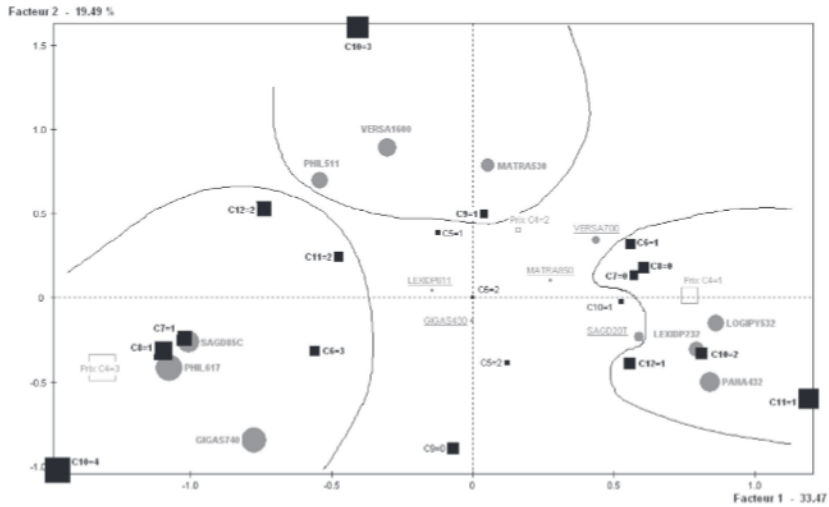


Figure 4.3 – Représentation simultanée standard des modèles de téléphone et de leurs attributs dans le plan des deux premiers facteurs

Or, dans cet exemple, la qualité des projections est globalement moyenne (valeurs des cosinus carré dans les tableaux 4.13 et 4.14) et une analyse directe effectuée à partir de la carte perceptuelle peut conduire à des erreurs d'interprétation. Ainsi, sur l'axe 1, les modalités C11-2 et C12-2 ne seraient probablement pas associées avec les modalités C7-1, C8-1 et C10-4. Intuitivement, elles seraient plutôt rapprochées du groupe PHIL 511 / VERSA 1600 et non de GIGAS 740, PHIL 617 et SAGBD 85C dont ils semblent *visuellement* plus éloignés. Par ailleurs, sur le pôle opposé, le modèle SAGD 20T serait intuitivement associé au groupe des trois appareils LEXI DP232, LOGI PY232 et PANA 432, alors qu'il n'est pas significatif sur cet axe ; de même, les modalités C6-1, C7-0 et C8-0 risquent de ne pas être associées aux modalités C10-2, C11-1 et C12-1 alors qu'elles doivent l'être (mais non C10-1 qui n'est pas ici significatif sur cet axe).

Ces « erreurs de perspective » dues aux distorsions de la projection peuvent entraîner des interprétations erronées. Rappelons que, pour interpréter un facteur à partir d'une représentation standard (où attributs et individus ont un rôle symétrique), il faut projeter sur l'axe considéré les points-colonnes (les modalités des attributs) ou les points-ligne (la position des individus) pour que les proximités au sein des éléments en ligne d'une part et les modalités en colonne d'autre part, deviennent immédiatement perceptibles (il ne faut surtout pas avoir une lecture « spatiale » du plan, directement fondée sur les distances entre points représentant les modalités ligne et colonne – *cf.* chapitre 3, section 2, § 2.4).

Pour situer les « individus » par rapport aux « attributs » – dans cet exemple, pour déterminer quelles sont les modalités des caractéristiques possédées ou non par les modèles de téléphone –, il faut utiliser le mode de projection barycentrique adapté

© Dunod – Toute reproduction non autorisée est un délit.

(ou *biplot*) dans lequel les modalités des variables sont projetées dans l'espace des individus (Figure 4.4). Ainsi, on peut affirmer que les modèles LOGY PY532, LEXI DP232 et PANA 432 sont caractérisés par les modalités C12=1, C11=1, C10=2, C8=0, C7=0 et C6=1, c'est-à-dire qu'ils ont l'agrément le plus faible, la portée et le nombre de services le plus réduits, ne possèdent ni la fonction baby-sitting ni l'éclairage de l'écran et, enfin, présentent une faible durée d'autonomie. À l'opposé, les modèles PHIL 617, SAD 85C et GIGA S740 disposent de l'équipement maximal sur les mêmes attributs associés à un meilleur agrément.

Comme pour l'analyse des correspondances d'un tableau croisé, cet exemple montre comment l'analyse des correspondances multiples permet d'analyser et de visualiser de manière synthétique, et avec efficacité, un tableau individus - variables, à la condition de prendre les précautions qui s'imposent dans la lecture et la projection des résultats.

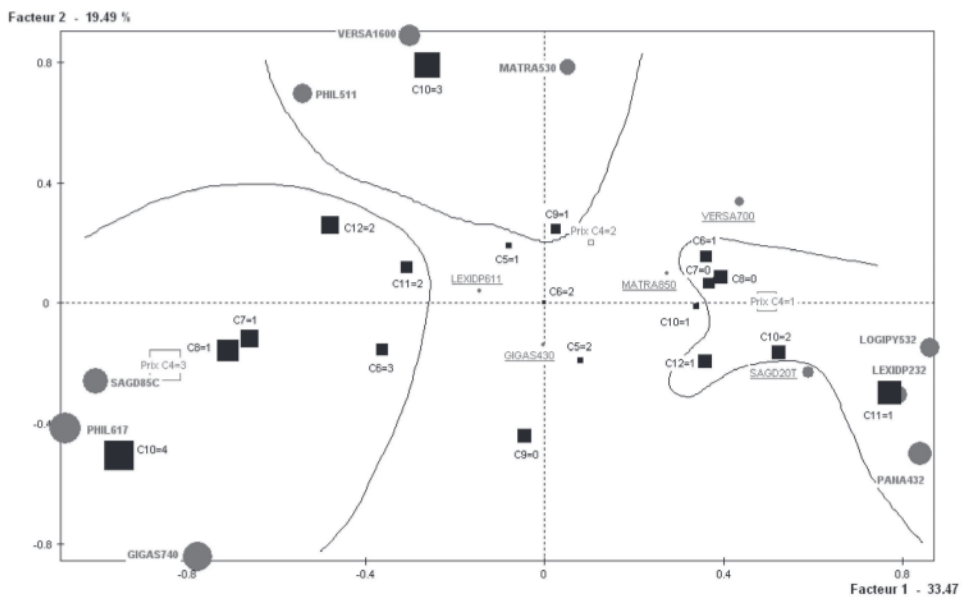


Figure 4.4 – Représentation des attributs dans l'espace des produits pour les facteurs 1 x 2 (projection barycentrique ou « biplot »)

L'ESSENTIEL

■ *Les objectifs d'une ACM*

L'analyse factorielle des correspondances multiples (ACM) a les mêmes fonctions qu'une analyse factorielle métrique, mais s'effectue sur des données non-métriques nominales, ce qui en fait un cas particulier de l'ACP. À la différence d'une AFC qui traite deux variables non métriques (*cf.* chapitre 3), l'ACM exploite un classique tableau de données, présentant les observations en ligne et les différentes variables en colonne. Mais toutes les variables devant être non-métriques, les variables métriques sont transformées en classes pour être exploitables. Le logiciel construit les tableaux disjonctifs et de Burt sur lesquels est appliqué le modèle mathématique de l'analyse des composantes principales.

L'ACM permet d'étudier et de visualiser les relations de voisinage (proximité-éloignement ou attraction-répulsion) qui existent :
- entre toutes les observations en ligne ;
- entre toutes les modalités en colonne ;
- et simultanément : entre toutes les observations en ligne et les modalités en colonne.

■ *Le déroulement d'une ACM*

L'analyse des résultats se déroule en plusieurs phases et de manière comparable à une AFC :

- identifier les axes significatifs à partir du tableau des valeurs propres ;
- identifier les modalités significativement liées à chaque axe en examinant leurs valeurs-tests, puis interpréter le contenu de ces axes conjointement avec le tableau des contributions absolues et celui des coordonnées ;
- vérifier la capacité du modèle retenu (choix du nombre d'axes significatifs) à prendre en charge la variance du maximum de modalités et évaluer sa validité globale (fraction de l'inertie totale du tableau de contingence prise en charge) avec le tableau des contributions relatives ;
- synthétiser les relations de voisinage entre modalités des variables et observations avec les représentations graphiques qui peuvent venir en complément de l'analyse indispensable des données chiffrées, en respectant les règles de lecture recommandées pour éviter toute interprétation erronée.

© Dunod – Toute reproduction non autorisée est un délit.

Chapitre

5

Comment construire et valider une typologie

OBJECTIFS

- Savoir formuler les objectifs de la typologie et déterminer les critères sur la base desquels les observations (individus ou objets) seront regroupées.
- Savoir construire les profils de comparaison pertinents en regard des objectifs et choisir l'algorithme de classification adapté.
- Passer du résultat statistique à la compréhension et à l'analyse des groupes retenus en utilisant les outils statistiques appropriés pour valider la typologie construite et l'utiliser dans une logique managériale ou décisionnelle.

SOMMAIRE

SECTION 1 Phase préparatoire : construire le profil de similarité

SECTION 2 Phase opératoire : obtenir différentes partitions

SECTION 3 Phase confirmatoire : valider une typologie

Les méthodes de classification ont pour objectif de regrouper dans des classes[1] homogènes des individus ou des objets (services, produits ou marques)[2], en fonction de leur proximité ou de leur ressemblance à partir d'un ensemble de caractéristiques les qualifiant. Les observations sont donc comparées sur la base d'un profil composé de variables (métriques ou non) qui, toutes, doivent être pertinentes en regard des objectifs de la typologie à construire : c'est une première contrainte. Par ailleurs, l'usage de ces méthodes impose de respecter trois principes, ce qui constitue une seconde contrainte :

– principe d'*exclusivité* : toutes les observations doivent être affectées à un groupe et à un seul à la fois (mais on peut isoler un groupe d'« inclassables ») ;
– principe d'*homogénéité* : chaque groupe doit être composé d'observations qui présentent un profil voisin (faible variabilité intragroupe) ; et tous les groupes doivent être aussi dissemblables que possible (forte variabilité intergroupe) ;
– principe de *parcimonie* : le maximum de la diversité et de la dispersion des observations doit être résumé en un minimum de classes.

Il existe une grande diversité d'algorithmes pour construire des typologies ; on peut les regrouper en deux classes :

– Les **méthodes hiérarchiques** permettent de construire une partition en établissant des liens hiérarchiques entre les groupes. Elles procèdent soit par divisions successives (de l'échantillon global vers chaque observation individuelle), soit par agrégations successives (de l'observation individuelle vers l'échantillon global). Ces liens sont visualisés par un « arbre de classification » appelé *dendrogramme*[3] (illustration au tableau 5.5). Le nombre de classes est déterminé *a posteriori*, après l'analyse des résultats observés. Cette méthode est utilisée quand le nombre de classes est inconnu ou incertain et quand le nombre d'observations est réduit : l'analyse du dendrogramme conduit à identifier le nombre de classes pertinentes pour rendre compte de la diversité de l'échantillon étudié.
– Les **méthodes non hiérarchiques** fournissent des classes indépendantes construites par agrégation directe des observations dont le profil est voisin. Cependant, avant d'effectuer les calculs, elles nécessitent de fixer *a priori* le nombre de classes. C'est la méthode préconisée pour étudier de gros échantillons.

Dans la pratique, le nombre de classes nécessaire pour structurer efficacement les observations est souvent inconnu. Aussi, méthodes hiérarchiques et non-hiérarchiques sont associées pour construire une typologie : la première (appliquée sur un sous-ensemble d'observations choisies aléatoirement si la taille de l'échantillon à

1. Dans cet ouvrage on considère « classe » et « groupe » comme des termes synonymes.

2. On peut même les utiliser pour classer les variables, mais il faut leur préférer les méthodes d'analyse factorielle décrites dans le chapitre précédent.

3. On rencontre aussi parfois le terme 'dendogramme' chez certains auteurs et dans certains algorithmes ; l'étymologie fait cependant pencher pour dendrogramme, du grec « dendro » = arbre et « graphein » = dessiner.

traiter est très importante) permet de formuler plusieurs hypothèses de classification qui sont ensuite testées puis optimisées avec la seconde.

On parle quelquefois de « classification automatique » pour désigner ces techniques. En fait, en dehors des calculs proprement dits, il n'en est rien. Comme l'application de tout algorithme de classification fournit une partition - un découpage des observations réparties entre différentes classes - il faut appliquer une démarche rigoureuse pour construire une typologie opérationnelle et valide. Cette validité dépend, non de l'algorithme utilisé, mais d'abord des critères choisis en amont des calculs pour construire les profils de comparaison, puis des moyens mis en œuvre après les calculs pour vérifier la qualité des résultats obtenus et la pertinence de la classification retenue. Cette démarche se décompose en trois phases :

1. *Une phase préparatoire pour construire le profil de similarité*
   - Sélection des variables devant composer le profil
   - Choix d'une mesure et calcul de la similarité entre les profils
2. *Une phase opératoire concernant l'effectuation des calculs*
   - Choix et application d'un algorithme d'agrégation
   - Analyse des paramètres statistiques pour comparer les partitions obtenues
3. *Une phase confirmatoire pour comparer et tester les partitions obtenues*
   - Utilisation de divers outils statistiques de comparaison et de validation
   - Dégager la typologie optimale en regard des objectifs décisionnels

Cas d'illustration

Une ONG au bénéfice de la recherche pour les maladies orphelines

Une ONG collectant des fonds pour la recherche médicale au bénéfice de certaines maladies orphelines, « reconnue d'utilité publique » et habilitée à recevoir des fonds du public (dons, legs, donations, permettant à ses donateurs de bénéficier d'une déduction fiscale) constate une érosion, certes lente mais très régulière, à la fois des sommes recueillies et du nombre de donateurs. Cette érosion peut s'expliquer par la disparition d'une fraction de ses donateurs historiques (dont la moyenne d'âge est supérieure à 70 ans) qui n'est pas compensée par le recrutement de plus jeunes, mais ce n'est probablement pas la seule cause. En effet, on constate d'importantes disparités selon les régions et, dans une même région, selon les secteurs géographiques (correspondant historiquement à des bassins de population homogène).

Une étude-test est conduite sur une région découpée en 47 secteurs pour formuler un diagnostic avant d'imaginer les solutions adaptées pour enrayer cette érosion,

☞

© Dunod – Toute reproduction non autorisée est un délit.

sinon améliorer les résultats dans le futur. Les donateurs étant identifiables sur la base de leur mode de paiement, on dispose annuellement d'informations concernant l'âge, le sexe, le secteur de résidence et le don : montant, périodicité et régularité. Connaissant le lieu de résidence, il est donc possible d'agréger certaines de ces données par secteurs géographiques.

Sur la base de ces agrégats, l'objectif est de savoir s'il existe des structures d'évolution du don (des « comportements sectoriels ») différenciables selon les secteurs géographiques couvrant les années 2012 à 2015 et qui pourraient aider à comprendre certaines raisons de l'érosion de la collecte – diminution du nombre de donateurs, diminution du don moyen, voire les deux combinés – afin de proposer aux acteurs locaux des orientations d'action adaptées à leur contexte local. Le tableau ci-dessous fournit quelques exemples des données disponibles :

Secteur :	Donateurs Total 2012	Montant 2012 Total en €	Don moyen 2012 en €	/// ///	Donateurs Total 2014	Montant 2014 Total en €	Don moyen 2014 en €	/// ///
secteur 7	498	86853	174	///	490	90 650	185	///
secteur 10	188	20920	111	///	189	22 869	121	///
secteur 26	285	25650	90	///	225	22 500	100	///
secteur 30	224	14112	63	///	180	12 420	69	///
secteur 42	267	33482	125	///	226	30 736	136	///
secteur 43	125	29147	233	///	84	17 136	204	///
///	///	///	///	///	///	///	///	///

Section 1 PHASE PRÉPARATOIRE : CONSTRUIRE LE PROFIL DE SIMILARITÉ

1 Préliminaires à l'élaboration d'un profil de comparaison : quel est l'objet de l'étude ?

Selon la nature – objective ou subjective – des critères utilisés pour comparer les observations, la typologie permet de traiter différents types de problèmes, à l'exemple des problématiques rencontrées en marketing (*cf.* tableau 5.1).

1.1 Étudier la similarité entre les individus (des clients ou des prospects)

La similarité entre les individus repose sur l'analyse de leurs attentes ou de leurs attitudes. C'est l'approche qui permet à l'entreprise d'être le plus en phase possible avec le marché. La typologie permet d'isoler des groupes spécifiques de prospects :

– pour analyser les offres les mieux adaptées aux exigences des clientèles et apprécier les positions concurrentielles ;
– pour identifier les caractéristiques que les offres devraient posséder pour satisfaire les différents groupes de prospects ;
– pour déterminer quelle stratégie de ciblage adopter : définir une offre spécifique par segment ou envisager la possibilité d'une offre globale susceptible de satisfaire le plus grand nombre.

La similarité repose sur les habitudes et les contextes de consommation des prospects. Cette approche permet de mieux comprendre le consommateur dans son rapport aux produits et d'identifier la structure d'un marché telle qu'elle est perçue par la clientèle.

> **Exemple**. L'analyse de la mixité de la consommation de produits laitiers frais a mis en évidence qu'il existait plusieurs structures de consommation qui correspondaient à des fonctions et des représentations très spécifiques de ces produits. Ce constat a permis de proposer des offres nouvelles en affirmant leur positionnement comme « dessert » pour définir leur statut sans ambiguïté et éviter toute confusion sur le contexte de consommation.

Tableau 5.1 – Exemples de problématiques marketing traitées avec la typologie

Objet de l'analyse :	Critères d'analyse :	
	objectifs	subjectifs
Individus	Habitudes d'usage : *Fonctionnalité des offres*	Attentes, attitudes, etc. : *Ciblage*
Produits et marques	Usage /consommation : *Situation concurrentielle*	Image, représentations : *Positionnement*

1.2 Étudier la similarité entre des objets (des produits ou des marques)

La similarité entre les offres ou les marques repose sur des critères de consommation ou d'usage : la typologie regroupe dans une même classe tout ce qui est perçu comme étant semblable. Cette approche structure donc la concurrence en identifiant les produits qui sont substituables et directement concurrents ou les marques utilisées conjointement (appartenance à un même groupe). Les produits ou les marques qui appartiennent à des groupes différents sont donc mutuellement complémentaires et non directement concurrents : ils couvrent des fonctions ou des besoins différents.

> **Exemple**. Si une huile d'arachide et une huile d'olive sont affectées à deux groupes d'usage différents, cela indique qu'on peut les trouver simultanément dans un même foyer et révèle une mixité d'usage (sans doute pour des utilisations culinaires différentes) : elles

© Dunod – Toute reproduction non autorisée est un délit.

> sont complémentaires. Le fait qu'un beurre et qu'une margarine soient dans un même groupe indique que leurs usages sont interchangeables : il y a donc substituabilité entre ces produits, l'utilisateur faisant le choix de l'un ou de l'autre, mais non des deux à la fois.

La similarité repose sur des critères d'image ou sur les représentations des prospects : cette approche permet d'analyser le positionnement. Les offres appartenant au même type partagent subjectivement les mêmes qualités et sont donc considérées comme équivalentes : leur positionnement est perçu identique. Les offres qui appartiennent à des groupes différents sont bien différenciées et possèdent un positionnement spécifique. L'analyse des classes obtenues permet de vérifier si le marché étudié est structuré autour d'un positionnement bien marqué (marques ou produits clairement identifiés au sein de types très différenciés les uns des autres) ou s'il est plutôt flou, voire inexistant (types peu différenciés, de structure voisine).

2 Sélectionner les variables pour construire le profil de comparaison

Le nombre de variables qui constituent le profil de comparaison a une incidence sur la « classifiabilité » des observations et la facilité à construire aisément une typologie. Deux observations sont identiques si elles se ressemblent sur tous les critères analysés. Mais plus le nombre de critères de comparaison est élevé, plus il devient difficile de trouver deux observations strictement identiques. Il convient donc de construire un profil de comparaison qui soit aussi économique que possible, tout en offrant un contenu suffisant, pertinent et cohérent en regard des objectifs de l'étude.

2.1 Définir le contenu et la nature des critères de comparaison

Idéalement, le profil de comparaison ne devrait être composé que des seuls critères strictement nécessaires à la classification. Par ailleurs, la nature de ces critères doit être homogène et chacun doit apporter une information utile, complémentaire des autres pour affiner la comparaison, en évitant toute redondance informative inutile.

Pertinence des informations en regard des objectifs

Les produits, les marques et tout objet en général, tout comme les individus, peuvent être comparés sur des critères qui décrivent leurs caractéristiques intrinsèques (critères « objectifs ») ou sur les évaluations dont ils ont pu faire l'objet (critères « subjectifs »).

Exemple. Une annonce publicitaire peut être décrite en fonction des figures de style qui la composent : métaphore, litote, oxymore, etc. (critères objectifs) ; elle peut l'être aussi à partir des jugements recueillis dans un pré-test auprès de la cible, sur sa qualité esthétique, sa crédibilité, etc. (critères subjectifs). Dans le domaine de l'assurance, un assuré peut être décrit par le nombre et la nature des sinistres déclarés au cours des trois dernières années (critères objectifs) ; il peut l'être aussi à partir de ses attentes à l'égard d'une compagnie d'assurances (critères subjectifs).

La sélection des critères composant le profil dépend des objectifs de la typologie, tels que présentés dans le tableau 5.1. Si le profil de comparaison intègre à la fois des critères objectifs et subjectifs, il faut vérifier si ce mélange est bien cohérent et justifié. Ainsi, associer dans une même analyse des critères techniques caractérisant un produit et des critères d'image de marque est difficilement justifiable pour classer les offres, mais peut l'être pour classer les individus à partir de leur perception des produits et des marques.

Le nombre de données exploitées concerne le jugement ou l'opinion d'individus recueillis sur des objets. Or cette information peut servir à classer tant les individus que les objets. Pour éviter toute erreur dans le traitement, il faut bien préciser, des objets ou des individus, ce qui fait l'objet de la classification pour sélectionner et organiser les données en conséquence.

Exemples. Un protocole standardisé de pré-test publicitaire a permis d'agréger des fichiers de données sur un échantillon d'annonces publicitaires ; toutes ces données sont de nature subjective (avis des prospects). Avec les mêmes informations, ce fichier peut servir à classer les annonces ou les prospects. Selon l'option, la matrice des données traitée ne sera pas la même : agrégation des réponses des individus pour classer les annonces, ou agrégation des évaluations portées sur les annonces pour classer les individus.

Dans le cas ONG présenté, le fichier des données initiales contient par année et pour chaque donateur, le montant et le mode de son versement, sa localisation, etc. L'objet de l'étude n'est pas le donateur, mais le secteur géographique pour savoir si les secteurs ont des profils d'évolution différenciés : les données individuelles sont donc agrégées pour fournir par secteur les informations annualisées nécessaires, tous donateurs confondus. Par ailleurs, l'objet de l'étude concerne l'évolution de certaines de ces informations au cours des années : montant du don, nombre de donateurs et don moyen. Il faut donc créer une variable qui en rende compte, telle que le ratio x_t/x_{t-1}, à l'exemple de la variable « variation du nombre de donateurs entre 2014 et 2015 » (« *var donat 15/14* »).

Utilité et complémentarité informatives des critères

Pour construire le profil de comparaison, la tentation est forte d'utiliser tous les critères de même nature liés aux objectifs, pour ne rien manquer. Mais dans nombre de secteurs économiques, une typologie n'est considérée comme opérationnelle que si elle comporte un nombre limité de classes. Il est donc nécessaire d'être raisonnable dans le choix des critères et de leur nombre, sans trop simplifier (une variable dichotomique ne fournit que deux groupes) et sans trop compliquer : la dispersion

© Dunod – Toute reproduction non autorisée est un délit.

augmente avec la multiplication des critères et la longueur des échelles de mesure. Aucune typologie ne pourra rendre compte de manière économique d'une grande disparité (« variance ») des observations à classer.

> **Exemple**. Comparer des individus en retenant trois couleurs pour la chevelure (blond/châtain/noir) conduit à constituer trois groupes. Y ajouter la couleur des yeux (bleu/non bleu) double l'effectif. Y ajouter le genre (H/F) conduit à douze groupes potentiels… sur la base de trois critères simples !

Être raisonnable dans le choix des critères consiste à s'interroger sur leur contenu en regard des objectifs visés, pour ne retenir que les critères utiles, exclusivement liés aux objectifs. C'est aussi ne retenir que des critères complémentaires dont chacun apporte une information qui complète utilement les autres en apportant quelque chose de spécifique et d'important. Satisfaire ces contraintes pourrait se résumer autour de deux actions : la « chasse aux intrus » et la « chasse à la redondance ».

La chasse aux intrus consiste à rejeter tout critère qui est en décalage avec les contenus fondamentaux sur lesquels la typologie doit être construite. Le repérage et l'élimination des intrus résultent d'un examen attentif du contenu des items utilisables. Dans l'étude portant sur les sirops de fruits (*cf.* chapitre 4), par exemple, tous les critères utilisés sont liés à trois dimensions : les caractéristiques organoleptiques perçues, le niveau de gamme et le cadre de consommation. Parmi la douzaine d'items utilisés, l'un concerne la publicité (« fait une publicité gaie ») qui génère à lui seul un facteur. Cet item – imposé par l'agence publicitaire – est sans lien direct avec les dimensions à la base de l'analyse ; il doit être supprimé. Dans l'étude ONG, l'objectif étant de rechercher les secteurs géographiques qui présentent une évolution de leur générosité comparable sur les quatre années étudiées (évolution du montant et du nombre de donateurs), une variable comme l'âge n'est donc pas pertinente. Par contre, la localisation l'est puisqu'elle permet d'affecter chaque donateur à un secteur afin d'agréger les données.

La chasse à la redondance consiste à repérer tous les critères sémantiquement voisins en se demandant en quoi ils peuvent se différencier et en quoi la distinction se justifie dans le cadre de l'analyse à mener.

> **Deux exemples** illustrent concrètement la « *chasse à la redondance* ».
> Dans l'étude ONG, le montant total des dons recueillis par secteur résulte du produit entre le nombre de donateurs et le montant du don moyen ; on peut donc faire l'économie de la variable résultante qu'est le « montant total » par secteur ;
> Une étude sur le café comprend les deux items suivants : « *bien parfumé* » et « *riche en arôme* » ; on peut penser qu'ils sont redondants parce qu'ils expriment la même idée. Par contre, cet autre item : « *parfum suave* » sera dissocié des précédents car il précise la nature de l'arôme et cet aspect peut être différenciateur des cafés testés.

Comme pour la recherche des intrus, la redondance peut être réduite à partir de l'examen minutieux du contenu des items. Mais elle peut reposer sur une démarche statistique. L'analyse des corrélations entre critères métriques ou des associations

entre critères non métriques (χ^2) permet d'identifier les items liés. Quand ces critères sont nombreux, l'application d'analyses factorielles (des correspondances ou en composantes principales) permet de constituer des facteurs associant tous les critères voisins. Ces facteurs qui condensent l'information brute peuvent remplacer les variables initiales : chaque observation est alors qualifiée par de nouvelles valeurs issues du calcul des scores factoriels liés aux axes retenus. Dans l'étude sur le café, par exemple, l'ensemble des items est regroupé après analyse factorielle autour de quatre axes : la couleur, l'arôme, le goût et la consistance en bouche. La typologie a été construite sur la base des scores factoriels issus de ces quatre dimensions remplaçant la quinzaine d'items initiaux. Dans l'étude ONG, l'analyse en composantes principales (ACP) effectuée sur les six variables initiales (variation du nombre de donateurs et du don moyen par paires d'années entre 2012 et 2015) fournit deux facteurs significatifs ($\lambda > 1$). Mais le troisième facteur ($\lambda = 0{,}89$) est aussi retenu car il fournit une information spécifique et complémentaire aux deux premiers axes. Le tableau 5.2 montre les scores de quelques secteurs sur les trois facteurs qui servent de variables et sur lesquelles la typologie est construite.

Tableau 5.2 – Les trois facteurs significatifs obtenus avec l'ACP

Secteurs	Facteur 1	Facteur 2	Facteur 3
Secteur 1	1,310	1,185	0,096
Secteur 2	1,387	1,336	0,018
secteur 3	1,265	0,886	0,206
Secteur 4	1,244	0,920	0,171
secteur 5	1,339	1,315	0,078
…	…	…	…
Secteur 43	0,340	-3,580	-1,291
Secteur 44	-0,112	-3,164	-1,272
Secteur 45	0,242	-3,266	-1,319
Secteur 46	0,270	-3,455	-1,103
Secteur 47	0,667	-2,934	-1,447

2.2 Choisir un nombre économique de variables

Au-delà de la pertinence de son contenu, pour être opérationnel, le profil de comparaison ne devrait contenir que les seules informations statistiquement nécessaires à la typologie : les critères doivent permettre de bien discriminer les observations. Enfin, sur le plan informatif, ce profil ne devrait comporter que des critères complémentaires, portant ainsi une information structurellement équilibrée.

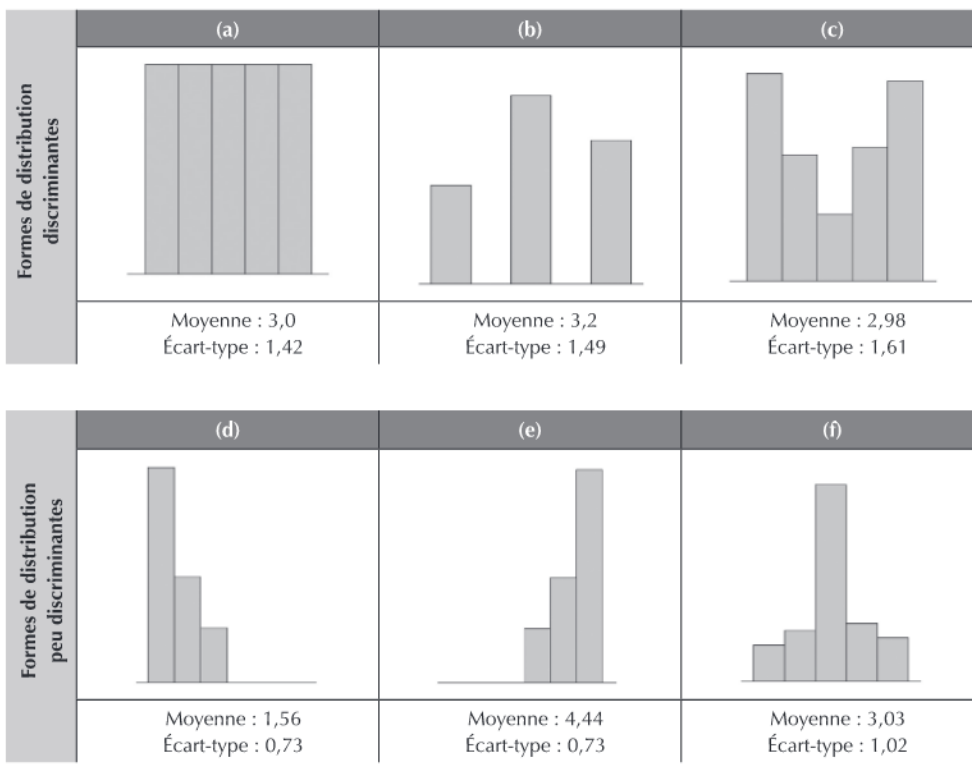
© Dunod – Toute reproduction non autorisée est un délit.

La capacité à discriminer les observations comme critère de sélection des variables

Une variable n'est intéressante pour comparer les observations que si elle permet de bien segmenter les observations. C'est une condition pour l'introduire dans le profil. Le pouvoir discriminant d'une variable est établi à partir d'une analyse statistique de la distribution de ses modalités pour une variable non-métrique ou de sa variance pour une variable métrique. Généralement, une faible dispersion traduit l'absence de différences sensibles entre les observations sur leurs caractères descriptifs ou un niveau de consensus élevé pour des items d'opinion. Dans ce cas, la variable n'est pas discriminante.

La figure 5.1 propose plusieurs exemples types de distribution. Un critère est d'autant plus discriminant que la répartition des observations entre ses différentes modalités ou valeurs est très étale (exemple 'a'). Il est d'autant moins discriminant que toutes les observations sont regroupées sur une modalité ou presque (d/e/f). Les distributions qui scindent les observations sur les modalités opposées (c) ou qui les répartissent sur quelques modalités bien spécifiées (b) sont aussi très discriminantes. L'analyse des écarts-type pour les variables métriques permet d'apprécier directement la valeur discriminante des items.

Figure 5.1 – Formes de distribution discriminant les observations



Recherche d'un profil structurellement homogène et équilibré

Disposer de variables indépendantes

Cet aspect complète la maîtrise de la redondance informative abordée plus haut. Le contenu des critères permet de déceler les redondances sémantiques, mais certaines variables peuvent être liées bien que leur signification soit très différente. Seule, une étude systématique des liens entre paires ou blocs de variables permet de relever ce niveau de redondance informative (coefficients de corrélation pour les variables métriques, coefficient de contingence pour les variables non métriques, analyses factorielles adaptées à la métrique des blocs de variables analysées).

> **Exemple**. Parmi les critères utilisés pour comparer une série d'annonces publicitaires en vue de les classer, il est apparu que deux variables métriques, *a priori* indépendantes, étaient en fait très liées : l'appréciation globale de la publicité et la compréhension subjective du message. Toute publicité qui n'était pas comprise assez rapidement recevait systématiquement une appréciation négative. Ce type de dépendance « psychologique » entre variables n'est pas rare. Dans cet exemple, la compréhension est prépondérante pour le prospect, mais l'appréciation en rend compte : à l'analyste de décider lequel de ces deux critères (ou les deux) intégrer dans le profil.

Donner le même poids à tous les critères

Plusieurs variables fortement liées présentes dans le profil conduit implicitement à donner un poids plus important à certains aspects de la comparaison qu'à d'autres. Ce choix est bien évidemment possible, mais il doit être explicité et justifié[1].

L'usage d'une méthode factorielle (ACP ou ACM) conduit à structurer ces liens au sein de facteurs homogènes quant à l'information prise en charge par chacun d'eux. Par ailleurs, les axes sont indépendants par construction (puisque orthogonaux, du moins dans un usage classique de ces modèles). Donc, chaque axe apporte une information originale et complémentaire par rapport aux autres ; ainsi, les observations sont comparées sur des variables agrégées indépendantes les unes des autres et pesant le même poids : les facteurs.

Neutraliser la longueur des métriques

L'éventuelle diversité des métriques utilisées induit des disparités dans le poids des variables composant le profil de comparaison. Plus la métrique est longue, plus elle génère de variance et cela entraîne une pondération artificielle des variables. Ce type de disparité existe dans l'exemple ONG : par secteur, le nombre de donateurs se limite à quelques centaines quand le montant recueilli oscille entre quelques dizaines et centaines de milliers d'Euros : cette variable à mécaniquement un poids en moyenne dix fois plus important en raison de son étendue.

1. L'estimation de la ressemblance entre profils reposant sur certains indices de distance (comme la distance euclidienne) peut être biaisée par l'existence de corrélations entre critères de comparaison.

© Dunod – Toute reproduction non autorisée est un délit.

Les procédures statistiques de *normalisation* ou de *standardisation* pour des variables métriques (*cf.* chapitre 1) permettent de résoudre ce type de problème : la variance des variables étant ramenée à l'unité, toutes présentent donc la même structure de dispersion. Les algorithmes de typologie proposent cette option avant d'effectuer les calculs. Cependant, si la dispersion est l'information qui justifie l'étude typologique, ces procédures de transformation ne doivent pas être utilisées.

> **Exemple**. Un certain nombre de juges ont évalué subjectivement des spots télévisuels sur une batterie de critères usuels dans les études publicitaires. Le même fichier de données peut être utilisé à deux fins : comparer les publicités (usage habituel) ou comparer les juges. Pour comparer les publicités entre elles, il faut supprimer un éventuel « effet juge » (certains notent systématiquement plus sévèrement que d'autres) ; la standardisation des évaluations par juge supprime cette disparité et leurs évaluations sont comparables. Mais comparer les juges entre eux oblige à tenir compte de leur plus ou moindre sévérité ; il faut donc travailler sur les évaluations brutes des spots qui conservent la dispersion initiale.

Dans pratiquement toutes les enquêtes internationales, pour rendre comparables les variables, la standardisation s'impose afin d'éliminer l'effet des particularités culturelles locales, d'autant plus qu'elles conduisent à utiliser des échelles d'amplitude différente selon les pays. Par ailleurs, les répondants ont souvent des comportements différents, certains n'utilisant qu'une partie de l'amplitude de l'échelle quand d'autres couvrent son intégralité.

Appliquer des pondérations directes

Indépendamment de ces procédures, certains algorithmes permettent de pondérer directement les critères inclus dans le profil de comparaison, en minorant ou en majorant leur rôle dans le calcul des similarités ou des distances entre les observations. Quelque peu empiriques, ces pondérations reposent sur une certaine subjectivité et leur usage implique une excellente maîtrise du domaine étudié, avec une grande prudence dans leur application. Cette subjectivité en fixe les limites et conduit à préférer, chaque fois que cela est possible, l'usage de méthodes statistiques pour élaborer un profil de comparaison qui soit homogène et équilibré, tant dans la nature que dans le poids accordé aux critères retenus.

3 Construire une mesure de la similitude entre les observations

L'affectation d'observations à un même type est déterminée par leur degré de ressemblance mesurée avec un indice qui définit objectivement leur proximité. Le choix de cet indice dépend de la nature des variables qui composent le profil de comparaison : indices de distance pour les variables métriques et indices de similarité pour les autres cas.

3.1 La mesure de la distance entre les observations (variables métriques)

Deux observations sont d'autant plus semblables que la distance qui les sépare est petite ; et, inversement, elles sont d'autant plus dissemblables que cette distance est élevée. Trois indices de distance sont parmi les plus utilisées (*cf.* tableau 5.3-A) :

– La *distance euclidienne* privilégie les grandes différences entre les observations au détriment des petites (puisque les écarts sont élevés au carré). De ce fait, son utilisation s'impose pour traiter un ensemble d'observations très hétérogènes. Mais elle est sensible aux corrélations entre variables et ne devrait être utilisée qu'avec des critères rendus indépendants.
– La *distance de Manhattan* (ou « *city block* ») privilégie les petits écarts entre les observations (puisque les écarts ne sont pas élevés au carré). Elle est adaptée pour structurer un ensemble d'observations dont l'hétérogénéité n'est pas très prononcée. Cette distance est aussi adaptée pour traiter des variables binaires.
– La *distance de Mahalanobis* est une mesure de distance pondérée, indépendante des corrélations entre variables. Son usage est recommandé pour comparer des profils dont toutes les variables ne sont pas indépendantes.

Il existe bien d'autres indices de distance proposés par les logiciels de traitement, à l'exemple de SPSS qui en offre une longue liste.

Tableau 5.3 – Les indices pour mesurer la distance ou la similarité entre les observations

A - Les différents indices de distance		
Types de distance	**Formulation**	**Caractéristiques**
Distance euclidienne	$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}$	Privilégie les grandes distances. Pour variables indépendantes.
Distance de Manhattan ou « City block »	$d_{ij} = \sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)$	Privilégie les petites distances. Applicable à des variables binaires.
Distance de Mahalanobis	$d_{ij} = \left(x_i - x_j\right)' S^{-1} \left(x_i - x_j\right)$	Recommandé quand les variables ne sont pas indépendantes.
Distance du Chi-Deux	$d_{ij} = \sqrt{\sum_{k=1}^{p} \frac{x_{..}}{x_{.k}} \left(\frac{x_{ik}}{x_{i.}} - \frac{x_{jk}}{x_{j.}}\right)^2}$	Recommandé pour traiter les variables nominales (comparaison sur modalités)

© Dunod – Toute reproduction non autorisée est un délit.

B - Les différents indices de similarité		
Types de similarité	**Formulation**	**Caractéristiques**
Indice de Sokal et Michener	(a + d) / (a + b + c + d)	Rôle équilibré entre toutes les coprésences et différences*
Indice de Rogers et Tanimoto	(a + d) / (a + d) + 2 (b + c)	Accent sur les différences
Indice de Jaccard	a / (a + b + c)	Accent sur les coprésences
Indice de Russel et Rao	a / (a + b + c + d)	Accent sur les coprésences
Indice de Sokal et Sneath	a / a + 2 (b + c)	Réduit l'effet des différences

*On peut démontrer que cet indice correspond à la distance de Manhattan

3.2 La mesure de la similarité entre les observations (variables non-métriques)

Les profils de comparaison peuvent aussi être composés de variables non métriques, ordinales ou nominales.

– *Cas des variables nominales*. Les données peuvent être multimodales (chaque variable est composée de plusieurs modalités) et traduites sous forme de comptage : l'indice du Chi-deux est bien adapté pour mesurer la distance entre les observations.
– *Cas de variables ordinales* – par exemple quand le profil est le résultat de choix ou de préférences émis par des individus, ou celui de classements et hiérarchies construits par l'analyste – la proximité entre deux profils est mesurée à partir des coefficients de corrélation de rang (de Spearman ou de Kendall)[1].
– *Cas des variables nominales de forme binaire* (liées au fait de posséder ou non un caractère donné). Ce cas donne lieu au calcul d'un indice de similarité entre deux observations ; il est fondé sur l'analyse de l'ensemble des attributs composant le profil de comparaison. Les divers indices (tableau 5.3-B) tiennent compte du nombre de coprésences (a : 1/1) d'attributs partagés par les deux, du type de différence (possédé par l'une et non par l'autre : b : 1/0 et c : 0/1) et des non-coïncidences (partagé par aucune des deux : d : 0/0).

La valeur de ces indices varie entre 0 (absence de similarité) et 1 (similarité totale). Mais sur les mêmes données, selon l'indice utilisé, l'intensité de la similarité calculée diffère : le tableau 5.3-B énumère ces mesures en ordre décroissant, des valeurs les plus élevées aux plus faibles. Le choix d'un indice de similarité dépend donc de l'importance accordée, à la présence simultanée ou à l'absence simultanée des caractères fondant la comparaison, telles que les formules l'expriment.

1. Voir chapitre 1. Si ces classements donnent lieu au calcul d'une « fonction-score » qui permet d'obtenir des vecteurs de rang liées aux paires comparées, on peut aussi utiliser une distance euclidienne.

Pour aller plus loin – Exemple de calcul des distances

Dans l'exemple « ONG » (*cf.* tableau 5.2), on peut calculer la distance entre les secteurs 4 et 44 sur les trois mesures adaptées aux données quantitatives et constater les différences de valeurs obtenues. La distance euclidienne donne :

$$\sqrt{(1,244-(-0,112))^2+(0,920-(-3,164))^2+(0,171-(-1,272))^2}=4,539$$

le carré de la distance euclidienne entre les mêmes secteurs est égale à :

$$(1,244-(-0,112))^2+(0,920-(-3,164))^2+(0,171-(-1,272))^2=20,600$$

alors que la distance de Manhattan donne :

$$(1,244-(-0,112))+(0,920-(-3,164))+(0,171-(-1,272))=6,883$$

Le calcul des indices des distances produit une matrice carrée et symétrique qui indique les distances ou les similarités entre toutes les observations deux par deux. Lorsque les caractéristiques de la dispersion des variables sont très disparates, pour équilibrer le rôle de chaque variable sur le profil des observations, il faut standardiser les variables avant le calcul des proximités (les scores *z*, par exemple, mais d'autres méthodes sont aussi disponibles).

Section 2 PHASE OPÉRATOIRE : OBTENIR DIFFÉRENTES PARTITIONS

Une fois les profils de comparaison construits et l'indice de similarité choisi, il faut réaliser la classification proprement dite à l'aide d'un algorithme approprié. Le choix d'une méthode dépend d'abord de la possibilité de fixer *a priori* le nombre de groupes. Si ce nombre est connu, une méthode non hiérarchique peut être utilisée d'emblée ; s'il est inconnu, une classification hiérarchique permet de guider ce choix. Mais dans la pratique, méthodes hiérarchiques et non hiérarchiques sont souvent associées dans un même protocole d'analyse, d'abord pour déterminer le nombre plausible de groupes (classification hiérarchique), puis pour optimiser la répartition des observations entre les groupes (méthode non hiérarchique).

Cependant, ces deux méthodes sont suffisamment différentes dans leur fonctionnement pour qu'on en traite séparément.

© Dunod – Toute reproduction non autorisée est un délit.

1 Effectuer une classification hiérarchique

La classification hiérarchique consiste à découper un ensemble d'observations à partir d'une analyse systématique des distances mesurées entre chaque paire (d_{ij}) de telle sorte que deux observations affectées à une même classe soient proches entre elles et éloignées d'autres observations affectées à une autre classe. Cette méthode construit :

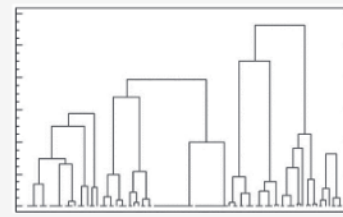
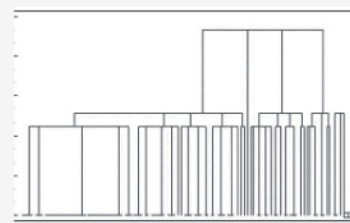
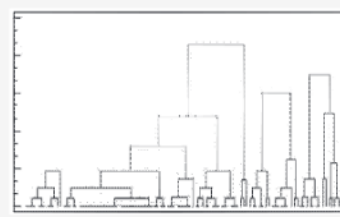
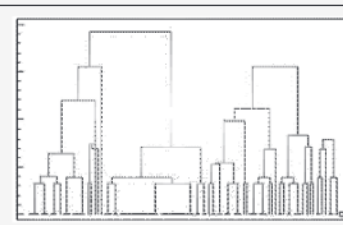
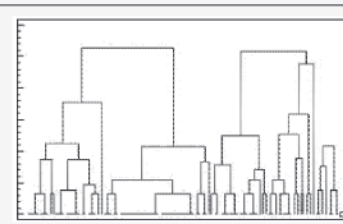
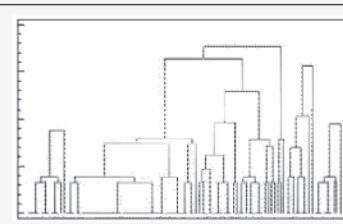
– soit une hiérarchie des rapprochements : la démarche, agrégative, part de l'observation individuelle (autant de groupes que d'observations initiales) pour atteindre par fusions successives la totalité de l'échantillon (toutes les observations rassemblées en un seul groupe) ; c'est le modèle de la *classification hiérarchique ascendante* et c'est le plus employé ;
– soit une hiérarchie de divisions, partant de la totalité de l'échantillon (un seul groupe) pour atteindre le stade de l'observation individuelle (autant de groupes que d'observations initiales) ; c'est le modèle de la *classification hiérarchique descendante*.

Quel que soit l'algorithme choisi, la procédure suit trois étapes : (1) choisir l'indice mesurant la distance entre les groupes pour définir les niveaux de genèse des groupes ; (2) identifier le nombre possible de classes à partir d'une analyse avec une méthode d'agrégation ou à l'aide de l'arbre de classification (« dendrogramme ») ; (3) repérer les observations marginales ou isolées pour décider soit de leur affectation à un groupe, soit de leur retrait du fichier analysé.

1.1 Choisir une méthode d'agrégation des observations

La genèse des groupes repose sur le calcul d'un indice qui mesure la distance séparant ces groupes et dont dépend la structure finale de la typologie (il importe de ne pas confondre cet indice de distance ou de fusion entre groupes avec ceux mesurant la distance entre les observations pour calculer leur degré de similarité – *cf.* section précédente § 3). Comme le choix de cet indice détermine le nombre de groupes générés et leur composition, il convient donc de le choisir en toute connaissance de cause car, selon l'indice utilisé, le nombre de groupes peut être différent. Les indices les plus usuels (avec leur nom anglais) sont décrits dans le tableau 5.4 qui présente aussi le type de diagramme qu'ils fournissent pour un même fichier de données. Comme le résultat obtenu – nombre et composition des groupes – dépend de l'indice utilisé, il importe donc de le choisir en connaissance de cause.

Tableau 5.4 – Les méthodes d'agrégation des observations

Type d'indice	Base de la distance entre groupes	Dendrogramme
Indice de Ward	L'homogénéité interne des groupes (variance intra-groupe) informant sur le degré d'homogénéité d'un groupe. Tend à générer des groupes d'effectifs équivalents	
Voisin le plus proche *Nearest neighbor* ou *Single linkage*	Deux observations entre lesquelles la distance est la plus réduite sont affectées à un même groupe. Tend à constituer des groupes d'assez grande variance car sensible aux premières affectations faites.	
Voisin le plus éloigné *Furthest neighbor* ou *Complete linkage*	La distance entre deux groupes est définie par la distance maximale entre les deux observations les plus éloignées d'un groupe à l'autre. Constitue des groupes très homogènes ; peu sensible aux points marginaux.	
Centre de gravité *Centroïd methode*	La distance entre centres de gravité (moyenne des observations par groupe). Implique que toutes les variables soient métriques.	
Moyenne de groupes *Group average* ou *Average linkage*	La distance moyenne de tous les points d'un groupe à l'ensemble des autres groupes.	
Distance médiane *Median*	Comme le précédent, mais fondé sur la médiane. Tend à fusionner les classes ayant une variance faible et à fournir une typologie dont les groupes ont un niveau de variance homogène.	

© Dunod – Toute reproduction non autorisée est un délit.

En gestion, c'est l'indice de Ward qui est le plus utilisé, car il tend à fournir des groupes de taille équilibrée, ce qui correspond bien aux exigences opérationnelles en matière de décision. Mais c'est aussi sa limite : il minimise la place d'observations originales qui, sans être marginales, peuvent être en nombre réduit dans l'échantillon. Cependant, divers travaux ont démontré que l'indice de Ward restait le plus performant pour classer des observations sur la base de la distance euclidienne.

La méthode d'agrégation de Ward se distingue de toutes les autres en ce sens qu'elle utilise une analyse de la variance approchée afin d'évaluer les distances entre classes. En résumé, cette méthode tente de minimiser la Somme des Carrés (SC) des distances entre tous les couples de classes pouvant être formés à chaque étape. L'usage de la méthode de Ward est justifié lorsque la « distance » entre les individus est le carré de la distance euclidienne[1]. Choisir de regrouper les deux observations les plus proches revient alors à choisir la paire de points dont l'agrégation entraîne la plus petite diminution de l'inertie du nuage. Le calcul des nouveaux indices entre la paire regroupée et les points restants revient alors à remplacer les deux points formant la paire par leur point moyen.

La plupart des logiciels de traitement de données offrent la possibilité de réaliser des classifications hiérarchiques et de construire automatiquement un dendrogramme en utilisant les différentes mesures de distances ou de similarité entre les observations et les différentes méthodes d'agrégation en classes.

1.2 Déterminer un nombre de classes avec le dendrogramme

L'arbre de classification ou « dendrogramme » est un outil utilisé pour déterminer le nombre de classes. C'est en fait une visualisation de l'histogramme des niveaux de fusion qui précise la position des objets étudiés : si l'arbre est horizontal, il présente en abscisse les observations et en ordonnée les valeurs de l'indice de séparation utilisé. La longueur des traits exprime la plus ou moins grande proximité entre les regroupements d'observations. Une faible longueur indique une forte proximité entre groupes d'observations autorisant à les fusionner dans une même classe ; une grande longueur indique une forte dissimilarité qui justifie de les séparer en des groupes distincts. Les différents dendrogrammes du tableau 5.4 illustrent la diversité des longueurs correspondant à des niveaux de fusion successifs.

La figure 5.2 reproduit le dendrogramme de la classification des 47 secteurs du cas ONG (obtenue sous SPAD avec la méthode du « voisin le plus proche ») sur la base des trois facteurs identifiés par l'analyse factorielle (voir tableau 5.2). Le choix du nombre de groupes repose sur l'analyse des variations de l'indice de fusion (2de colonne de la figure, avec « IND ». pour indice) où les valeurs élevées (surlignées) correspondent à des ruptures (« paliers ») qui définissent une partition (et se trouvent figurées par une plus grande longueur des branches). Il apparaît alors que cet

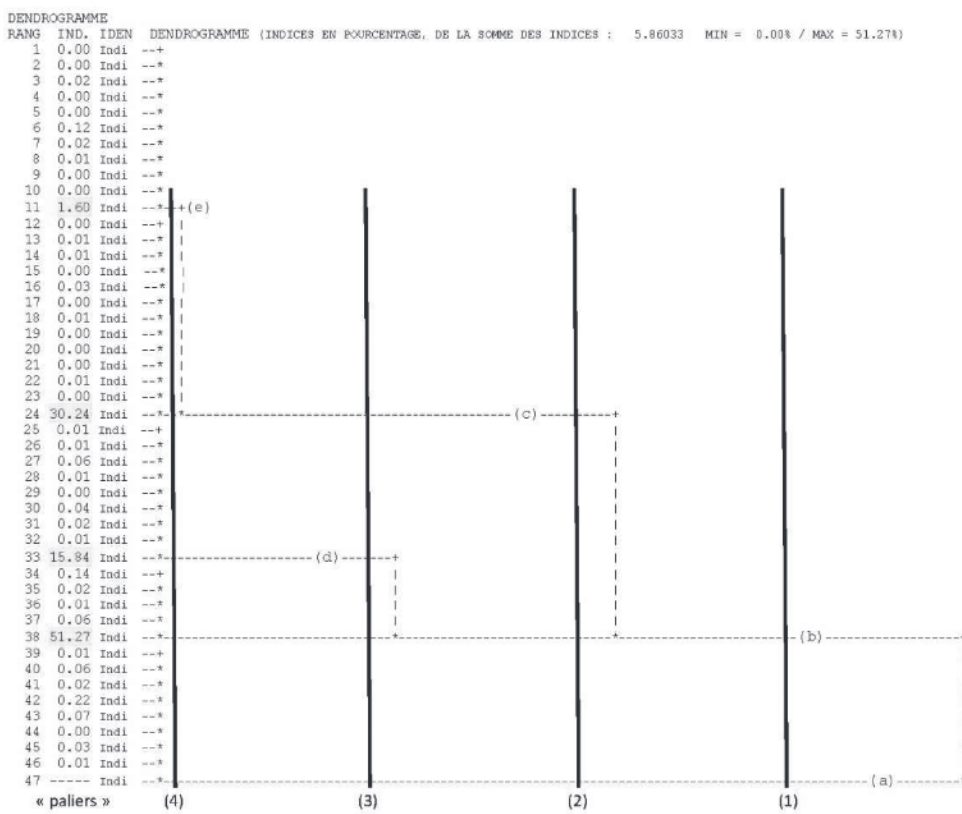
1. L'indice « carré de la distance euclidienne » doit être choisi quand une des méthodes d'agrégation CENTROID, MEDIAN ou WARD est sélectionnée.

ensemble peut être scindé en deux, trois, ou quatre groupes autonomes (voire cinq puisqu'il y a de nouveau une rupture au rang 11 même si l'indice de distance – 1,6 – est faible). Les niveaux de rupture sont surlignés et mis en correspondance avec les lettres dans le dendrogramme (*a*, *b*, *c*, *d* et *e*). La lecture des valeurs de l'indice permet d'identifier plus précisément ces sauts et juger des seuils qu'il est intéressant de retenir. Cette figure 5.2 reproduit l'histogramme des niveaux (*cf.* tableau 5.5) qui est à la source du dendrogramme et porte la même information.

Tous les éléments sont alors réunis pour entrevoir le nombre de classes plausibles :

– Le palier (1) correspond à une première variation importante (51,27) de l'indice de fusion utilisé (« voisin le plus proche ») : il définit un saut indiquant un premier niveau de division possible, ici en deux classes : la branche (a) composée des rangs 39 à 47 qui s'oppose à l'ensemble des autres rangs (branche 'b').
– Une seconde variation importante (15,84) apparaît au palier (2) où la branche (b) de l'arbre se scinde en deux : la branche (c) réunit les observations de rang 1 à 24 et la branche (b) les observations de rang 25 à 32 ; la troisième branche reste (a). À ce palier (2), il y a donc trois groupes possibles.

Figure 5.2 – Exemple de Dendogramme (cas « ONG »)



N.B. : les paliers ne sont pas dessinés par le logiciel mais par l'analyste

© Dunod – Toute reproduction non autorisée est un délit.

– Une troisième variation {palier (3) : (d) = 15,84} apparaît sur la branche (b) qui se divise en deux : elle est « amputée » des observations aux rangs 25 à 33 pour composer la branche (d) ; et (b) l'est des rangs 34 à 38. Quatre groupes se dessinent.
– Enfin, un ultime saut est envisageable au palier (4) {(e) : 1,60} qui décompose la branche (c) en deux ensembles (c) avec les rangs 12 à 24 et (e) avec les rangs 1 à 11. On note cependant que la longueur de la branche est très faible : la question se posera alors de décider quel modèle, entre 4 et 5 groupes, est le plus cohérent et opérationnel.

Le dendrogramme permet aussi de repérer les observations isolées ou marginales, qui constituent à elles seules une branche issue d'une variation importante de l'indice de fusion (il n'y a pas ce cas dans l'exemple ONG). Leur identification et l'examen de leurs caractéristiques permet de décider s'il faut les traiter comme groupes spécifiques ou s'il convient de les retirer du fichier de données. Ce choix n'est donc pas « automatique » mais résulte d'un travail de réflexion à mener.

L'analyse des « paliers » du dendrogramme fournit ainsi différentes hypothèses sur le nombre vraisemblable de groupes compte tenu de la similarité ou de la proximité entre toutes les observations étudiées. Ces hypothèses peuvent constituer la base des travaux ultérieurs pour rechercher la typologie la plus satisfaisante dans le cadre des objectifs de l'étude : chaque groupe doit alors être décrit pour identifier son contenu et pour juger de sa cohésion sur la base de sa variance interne (*cf.* Section 3 : phase confirmatoire).

2 Effectuer une classification non-hiérarchique

Une classification non hiérarchique ou méthode de réallocation, consiste à répartir un ensemble d'observations entre différents groupes dont le nombre a été défini auparavant. L'affectation repose sur l'analyse de la distance de chaque observation au centre de gravité (barycentre) de chacun des groupes constitués.

Parmi la diversité des techniques de réallocation qui existent, la méthode des K-means et celle des Nuées dynamiques, dont le processus est très voisin, sont les deux plus utilisées, particulièrement pour traiter de gros fichiers de données. Une limite cependant : le résultat final obtenu au terme des calculs est très dépendant de la configuration initiale engendrée par le choix du centre des groupes prédéfinis. Cette limite est dénommée « risque de minimum local ».

Après avoir choisi un indice de comparaison entre les observations (dans les mêmes conditions qu'en classification hiérarchique), l'analyste doit prendre deux décisions : (1) choisir le nombre de groupes et fixer leur centre de gravité, (2) puis fixer des règles pour l'arrêt des procédures de calcul.

Tableau 5.5 – Histogramme des indices de niveau (fichier « ONG »)

CLASSIFICATION HIERARCHIQUE (VOISIN LE PLUS PROCHE) SUR LES 3 PREMIERS AXES FACTORIELS
DESCRIPTION DES NOEUDS

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
48	13	17	2	2.00	0.00001	*
49	5	6	2	2.00	0.00001	*
50	15	19	2	2.00	0.00001	*
51	48	12	3	3.00	0.00002	*
52	3	4	2	2.00	0.00003	*
53	24	22	2	2.00	0.00003	*
54	51	18	4	4.00	0.00003	*
55	9	1	2	2.00	0.00005	*
56	23	14	2	2.00	0.00009	*
57	33	25	2	2.00	0.00011	*
58	2	49	3	3.00	0.00012	*
59	52	11	3	3.00	0.00015	*
60	16	21	2	2.00	0.00017	*
61	36	39	2	2.00	0.00018	*
62	10	55	3	3.00	0.00027	*
63	50	20	3	3.00	0.00036	*
64	63	56	5	5.00	0.00040	*
65	30	32	2	2.00	0.00040	*
66	59	8	4	4.00	0.00047	*
67	60	54	6	6.00	0.00048	*
68	27	31	2	2.00	0.00051	*
69	38	40	2	2.00	0.00053	*
70	67	53	8	8.00	0.00058	*
71	57	28	3	3.00	0.00058	*
72	43	46	2	2.00	0.00061	*
73	65	29	3	3.00	0.00067	*
74	35	37	2	2.00	0.00080	*
75	41	34	2	2.00	0.00097	*
76	66	7	5	5.00	0.00108	*
77	58	62	6	6.00	0.00110	*
78	72	45	3	3.00	0.00118	*
79	68	26	3	3.00	0.00124	*
80	70	64	13	13.00	0.00180	*
81	69	61	4	4.00	0.00184	*
82	79	71	6	6.00	0.00229	*
83	75	74	4	4.00	0.00363	*
84	82	73	9	9.00	0.00364	*
85	44	78	4	4.00	0.00376	*
86	81	42	5	5.00	0.00384	*
87	76	77	11	11.00	0.00675	*
88	85	47	5	5.00	0.00798	*
89	86	83	9	9.00	0.01299	*
90	80	87	24	24.00	0.09381	*** (5 classes)
91	88	84	14	14.00	0.92832	********** (4 classes)
92	91	90	38	38.00	1.77193	********** (3 classes)
93	89	92	47	47.00	3.00448	********** (2 classes)

SOMME DES INDICES DE NIVEAU = 5.86033

(Source SPAD)

© Dunod – Toute reproduction non autorisée est un délit.

2.1 Fixer le nombre de classes et les observations « étalons »

Avant d'effectuer les calculs, il faut fixer le nombre de classes voulu en s'appuyant, soit sur des informations déjà disponibles, soit sur l'expérience de l'analyste. Mais, le plus souvent, il résulte d'hypothèses fournies par une classification hiérarchique réalisée au préalable (quelquefois réalisée sur un sous-échantillon des observations à classer quand leur nombre total est élevé). Une fois ce nombre fixé, il faut fournir pour chaque classe l'observation initiale qui définit le centre de gravité de chacun des groupes. Plusieurs méthodes sont possibles :

– prendre les *k* premières observations correspondant au nombre *k* de groupes fixés ;
– sélectionner de manière aléatoire les *k* observations « représentant » ces *k* groupes ;
– déterminer de manière réfléchie les *k* observations initiales en fonction de la connaissance de leurs caractéristiques pour fixer la structure des groupes-types (par exemple, à partir de leurs coordonnées dans une structure factorielle).

Quelle que soit la méthode, ces observations qualifiées d'« étalons » servent de pivot à la constitution des groupes : toutes les autres observations sont affectées à l'un ou l'autre des groupes établis selon leur distance au centre de ces groupes.

Selon la manière de fixer les groupes initiaux, l'algorithme de calcul peut procéder de deux manières, en fonction de la commande à exécuter :

– *Le centre de gravité de chaque groupe est fixe* : les observations sont comparées systématiquement au centre de chaque groupe défini ; une observation est affectée au groupe dont elle est le moins éloigné. C'est la méthode la plus utilisée en cas d'affectation réfléchie et justifiée des étalons.
– *Le centre de gravité de chaque groupe est évolutif* (méthode des « centres mobiles ») : chaque observation est affectée au groupe dont elle est le plus proche, ce qui conduit à recalculer le centre de gravité du groupe. Toutes les observations sont de nouveau comparées aux centres de gravité recalculés, ce qui conduit certaines observations à migrer entre les groupes ; mais une observation n'est déplacée que si sa réaffectation améliore la cohésion des groupes. C'est la méthode habituellement retenue en cas d'affectation aléatoire des pivots.

2.2 Fixer certains paramètres nécessaires aux calculs

Selon l'algorithme proposé dans les logiciels utilisés, en plus du nombre de classes initiales, la procédure de calcul peut aussi imposer d'autres choix liés à la définition de la structure finale ou liés à l'affectation des observations aux classes. Ces choix sont assez empiriques et demandent de procéder à plusieurs essais.

Choix liés à l'identification de la structure finale (méthode des Nuées dynamiques)

– *Le critère de convergence*. La convergence, qui définit la stabilité des centres de gravité, est atteinte dès lors que les observations ne changent plus de groupe d'une itération à l'autre. On peut fixer un critère de convergence global : compris entre 0 et 1, il est lié à la distance minimale entre les centres de classe et exprime la fraction de déplacement en deçà de laquelle la position des classes est considérée comme stable, donnant une structure « définitive ».
– *Le nombre d'itérations maximum*. L'algorithme arrête les calculs une fois que le nombre autorisé d'itérations est atteint, même si la convergence n'est pas atteinte (si elle est atteinte avant épuisement du nombre d'itérations, l'algorithme s'arrête). Si la convergence n'est pas atteinte en moins d'une cinquantaine d'itérations, on peut s'interroger sur la « classifiabilité » des données (mais ce nombre peut être d'autant plus élevé que l'échantillon à traiter est de très grande taille).

Choix liés à l'affectation des observations (méthode des K-means)

– *La distance minimum entre classes* : c'est la distance entre les centres de gravité de deux groupes en deçà de laquelle ils seront fusionnés.
– *La distance maximum au centre d'une classe* : c'est la distance d'une observation au centre de gravité d'un groupe à partir de laquelle cette observation peut constituer le noyau d'un nouveau groupe. C'est donc le « rayon » d'un groupe, lequel permet aussi d'identifier les observations marginales (dont la distance à un groupe quelconque est supérieure à cette valeur).

Les méthodes de réallocation fournissent une solution le plus souvent satisfaisante sur le plan statistique, mais c'est une solution parmi d'autres envisageables et il n'est pas possible de démontrer que c'est la meilleure. Comme les solutions fournies ne sont pas stables, mais dépendantes du choix initial des étalons (risque de minimum local), cela contraint donc à effectuer plusieurs essais en modifiant les étalons. Une analyse comparée de la structure des groupes ainsi obtenus permet de repérer l'association systématique de certaines observations dans un même groupe, constituant ainsi la base d'une structure stable. Ces convergences, parfois appelées « *formes fortes* », peuvent constituer la base de groupes-étalons sur lesquels vont venir se greffer les autres observations.

Le risque de « minimum local » peut être réduit en effectuant une classification hiérarchique sur les observations à traiter (ou sur un sous-échantillon extrait aléatoirement si le fichier est de grande taille). Il a été démontré[1] que pour générer la configuration initiale nécessaire aux méthodes de réallocation, le choix des étalons reposant sur les résultats d'une classification hiérarchique produisait des résultats plus stables et de meilleure qualité qu'une définition aléatoire des étalons, d'autant plus qu'elle permet de fixer le nombre initial de groupes.

1. Everitt B., Landau S., Leese M. (2001) : *Cluster Analysis*, Oxford, Oxford University Press (4e édition).

© Dunod – Toute reproduction non autorisée est un délit.

Section 3

PHASE CONFIRMATOIRE : VALIDER UNE TYPOLOGIE

Quel que soit le type d'algorithmes utilisé (méthodes hiérarchiques ou non), il importe de vérifier la validité et la pertinence des résultats typologiques obtenus. Mais, bien qu'elle s'appuie sur des indicateurs statistiques, cette validation reste assez empirique et se déroule en deux étapes. La première consiste à déterminer la structure des classes qui rend le mieux compte, et de manière cohérente, des ressemblances et des différences entre les observations ; elle repose sur l'analyse du degré de convergence entre plusieurs solutions typologiques testées. La seconde consiste à décrire et comparer chacun des types pour cerner leurs différences de contenu et juger tant de leur pertinence que de leur opérationnalité.

1 Choisir une structure typologique pertinente

Comme tout algorithme fournit une partition des observations analysées, ce serait une erreur de considérer *a priori* qu'une structure issue des calculs est nécessairement satisfaisante. Il faut d'abord tester la *validité interne* du modèle avant de l'utiliser pour prendre des décisions. Tester la validité interne consiste à vérifier la stabilité des groupes obtenus et à étudier leur degré d'homogénéité.

1.1 Analyser la stabilité des groupes

La validité d'une typologie peut être éprouvée en analysant le degré de convergence des résultats issus de différents essais de classification. Ces essais peuvent être faits de plusieurs manières, d'ailleurs combinables entre elles :

– en modifiant systématiquement certains paramètres : les indices de similarité entre profils et les indices de fusion entre groupes ;
– en associant différentes méthodes : hiérarchique et non hiérarchique ;
– en sélectionnant des « étalons » différents pour constituer les groupes de base dans une classification non hiérarchique ;
– en travaillant alternativement sur des sous-ensembles de variables dont le contenu est comparable.

Certains logiciels proposent un protocole qui associe directement une analyse hiérarchique pour examiner d'abord le nombre de groupes plausibles, suivie d'une méthode non hiérarchique (Nuées dynamiques ou méthode des K-means) pour optimiser l'affectation des observations entre les classes construites. Ce type d'enchaînement, assez classique, est facile à mettre en œuvre quels que soient les logiciels d'analyse des données utilisés.

Les observations étant toutes affectées aux groupes constitués, la comparaison des différentes solutions obtenues – en faisant varier les paramètres, les variables ou les méthodes – permet de voir s'il existe des regroupements systématiques entre les mêmes observations. L'existence de tels « noyaux » est un indicateur concret du degré de classifiabilité des données.

Si le nombre d'observations est assez important, il est possible de scinder l'ensemble en deux sous-échantillons (voire trois) de taille comparable, mais constitués de manière aléatoire. Les soumettre à la même procédure typologique permet de vérifier la convergence entre les structures obtenues. Si, pour chacune d'elles, les centres de groupes (barycentres) sont comparables ou si chaque groupe est significativement décrit par les mêmes variables, on peut conclure à la validité du modèle dégagé.

D'autres indicateurs statistiques, comme la variance interclasse (la plus petite possible) ou la distance entre les centres de gravité de chaque groupe (la plus grande possible), contribuent à l'analyse de la validité en comparant l'évolution de ces valeurs sur les différentes classifications examinées.

1.2 Analyser l'homogénéité des groupes

L'homogénéité d'un groupe, sa cohérence interne, peut être analysée à partir de la variance des observations qui le composent. Plus les observations sont concentrées autour de leur centre de gravité, plus le groupe est homogène (faible variance interne ou « intra-classe »). Dans le même esprit, on peut analyser la plus grande distance entre deux observations au sein d'un même groupe pour apprécier sa cohésion interne.

Selon la nature des variables, deux indicateurs peuvent être utilisés :

– *sur des données métriques* : le rapport des variances intragroupe/intergroupe doit être aussi faible que possible ;
– *quelle que soit la nature des variables* : la distance des observations au centre de gravité du groupe. Plus il y a d'observations excentrées, moins le groupe est homogène. Le constat d'une observation très excentrée conduit à s'interroger sur les raisons de sa distance ; il permet aussi d'identifier des observations marginales (pour décider de leur éventuel maintien ou retrait de l'analyse).

L'application systématique de telles analyses sur chaque groupe et la comparaison qui s'ensuit conduit à évaluer la cohérence interne de chaque solution testée. Ce premier travail contribue à identifier le modèle typologique le plus structuré et susceptible d'être le plus valide et opérationnel. Encore faut-il qu'il soit interprétable pour être définitivement utilisable.

2 Déterminer la valeur opérationnelle d'une typologie

Après avoir conclu sur la validité interne du modèle retenu, il faut encore se prononcer sur sa validité externe et, pour le décideur, sur son opérationnalité. Chaque

© Dunod – Toute reproduction non autorisée est un délit.

groupe doit pouvoir être décrit de manière compréhensible et cohérente sur la base de toutes les variables liées au problème à traiter, autant celles qui ont servi à classer les observations (variables actives) que l'ensemble des autres (variables illustratives). Après avoir présenté les principes dans les paragraphes qui suivent, la démarche est illustrée dans le § 3 « illustration avec le cas ONG ».

2.1 Décrire chaque groupe

Chaque groupe est décrit sur la base de plusieurs données :

– D'abord, par les *coordonnées de son centre de gravité* (moyenne des valeurs des variables de classification calculée sur l'ensemble des membres du groupe). Mais cette lecture n'est pas toujours aisée, particulièrement en raison du caractère composite des variables de classification utilisées (des scores factoriels par exemple). Identifier l'observation « représentative » du groupe permet de surmonter cet obstacle. Il s'agit de repérer parmi celles qui sont le plus proche du centre de gravité, l'observation la plus typique comme « modèle » du groupe.
– Ensuite, par les valeurs des *variables actives* à la base de la typologie. Certains logiciels permettent de calculer pour chaque groupe un indicateur de significativité fondé sur le test de Student pour les variables métriques ou sur le Chi-deux pour les modalités des variables non métriques. Cette description, en s'appuyant sur les données statistiquement significatives, aide à l'interprétation du contenu des groupes.
– Enfin, par les valeurs des *variables illustratives* qui ne sont pas intervenues dans la classification mais qui sont pertinentes par rapport au problème posé.

De manière plus globale, l'usage de certains outils statistiques permet d'identifier les variables significativement liées à la typologie. Dans ce cas, la structure typologique est la variable nominale susceptible de rendre compte des écarts entre groupes sur les différentes variables étudiées. Cette pratique peut s'imposer si le logiciel de classification utilisé n'indique pas les variables significatives de chaque groupe. Plusieurs approches peuvent alors être combinées :

– Une analyse de variance effectuée sur les *variables actives* (de manière opportuniste puisqu'elles déterminent la typologie) permet de les hiérarchiser en fonction de leur degré de segmentation (séparant tous les groupes de manière significative ou ne marquant que des oppositions partielles). Cette ANOVA est fondée sur le test F pour les variables métriques ou sur le test de Kruskal-Wallis pour les variables ordinales. Pour les variables nominales, un test d'indépendance fondé sur le Chi-deux entre la distribution de chaque variable étudiée et la variable identifiant les groupes est effectué. Une analyse des correspondances permet aussi de situer les classes (ayant alors statut de variable illustrative) par rapport aux variables non métriques actives.
– Une analyse de variance sur les *variables illustratives* permet d'identifier celles qui sont affectées par la classification ; leur identification peut éclairer le contenu

des groupes ou leur signification. Dans tous ces cas, les comparaisons entre classes (« test des étendues multiples » ou « contrastes ») permettent d'identifier les variables, actives ou illustratives, significativement liées à la typologie.
– Une analyse discriminante permet d'identifier *globalement* quelles sont les variables liées de la manière la plus significative à la construction de la typologie.

2.2 Déterminer l'opérationnalité de la typologie

Il ne suffit pas qu'une typologie soit satisfaisante sur le plan statistique : il importe que ses résultats soient utiles pour aider le décideur dans sa réflexion. La recherche d'une typologie aussi fidèle que possible à la réalité du terrain, tout en restant opérationnelle, peut être envisagée sous deux angles complémentaires.

L'aide aux décisions stratégiques

L'utilité d'une typologie dépend de la nature du problème à résoudre. Dans le cas ONG, on recherche des structures de comportements différenciés pour suggérer des propositions d'action sur le terrain adaptées aux groupes de secteurs identifiés : des groupes d'effectifs faibles sont donc tout à fait recevables pour définir des actions spécifiques. Mais, en marketing par exemple, pour qu'un groupe soit intéressant à considérer dans le cadre d'une stratégie de ciblage, il faut qu'il corresponde à un potentiel, et donc un volume de clientèle, qui présente un réel intérêt économique.

> **Exemple**. Dans une étude sur les attitudes des ménages à l'égard de l'alimentation, six types ont été identifiés. Les attentes et les pratiques de l'un d'eux en faisaient un groupe-cible idéal pour les produits alimentaires industriels, mais ne représentait guère plus de 5 % des foyers, majoritairement composés d'une à deux personnes. Un autre type composé de ménages consommant peu de produits industriels, représentait en fait près du tiers des foyers. Après approfondissement, il est apparu que ce type, majoritairement composé de foyers de quatre personnes, constituait la source de revenus la plus importante pour de nombreuses références de l'entreprise.

Cependant, ce critère de volume peut conduire à biaiser la réalité en cherchant à obtenir des structures de taille équilibrée : la plupart des typologies réalisées comptent entre quatre et six classes composées de 15 % à 20 % des effectifs. Or, une « petite » classe peut être les prémices d'un groupe particulièrement typé, ayant ses caractéristiques propres : il peut présenter un intérêt pour diagnostiquer l'apparition de nouvelles attitudes, de tendances ou de comportements susceptibles de se développer dans un futur proche.

Les meilleures conditions pour obtenir une typologie optimale est de s'appuyer sur des hypothèses de structure fondée sur une connaissance du terrain. Dans ce contexte, l'application d'un algorithme de classification s'apparente à une analyse confirmatoire de ces hypothèses. L'absence de convergence entre ce qui est attendu ou a été prévu, et ce qui est obtenu après analyse implique alors de s'interroger pour

© Dunod – Toute reproduction non autorisée est un délit.

remettre en cause, soit le modèle hypothétique formulé, soit les bases qui ont servi à construire la typologie (les données et les variables utilisées).

La construction d'un modèle de représentation

Sur la base des variables du profil de comparaison qui a permis de construire et de valider la typologie, l'approche consiste à rechercher les fonctions mathématiques (combinaisons pondérées des variables) qui discriminent les groupes obtenus. Ainsi, dès lors que le profil d'une nouvelle observation est connu, ces fonctions permettent de *déterminer son appartenance* à l'un des groupes de la typologie constituée (cas du positionnement pour un objet et du ciblage en marketing ou du scoring en banque-assurances pour un individu). Quelques exemples d'application dans différents champs de la gestion illustrent l'intérêt de cette approche.

En marketing, dans la gestion des nouveaux produits

– *Cas d'un nouveau produit en test*. À partir d'une enquête réalisée pour qualifier le produit sur les composantes du profil, il est possible de l'affecter à un groupe préexistant. Compte tenu de ses caractéristiques perçues, on peut vérifier si son image est conforme au positionnement voulu ; et, s'il en diffère, analyser les raisons. Ainsi, l'affectation à un groupe permet d'identifier directement ses concurrents les plus proches, car tous les produits appartenant à un même groupe sont en situation de concurrence directe, quand l'appartenance à des groupes différents exprime une certaine complémentarité (*cf.* section 1, § 1.2).
– *Cas de l'arrivée d'un produit concurrent*. Selon la même procédure, on peut affecter tout nouveau concurrent au groupe dont il se rapproche le plus (ou découvrir qu'il est hors de la structure typologique de référence s'il est très novateur). Cela permet d'apprécier son positionnement et d'en évaluer le risque concurrentiel.

Dans le secteur bancaire

– Tout particulier qui envisage d'investir dans certains placements doit être classé en fonction du niveau de risque financier qu'il accepte. À cette fin, il est soumis à un questionnaire dont la diversité des items permet d'apprécier *a posteriori* son rapport au risque. À partir de son profil de réponse, il se trouve affecté à un groupe de niveau de risque qui contraint en principe le banquier à ne lui proposer que des produits conformes à ce niveau. C'est le principe du « scoring ». En cas d'une demande de prêt, le prêteur peut évaluer le risque qu'il prend d'être ou non facilement remboursé à partir d'une démarche similaire.

Une autre application de ces fonctions discriminantes consiste à vérifier la consistance de la structure typologique dégagée (*cf.* tableau 5.6). Elles sont utilisées pour réaffecter toutes les observations à leur groupe d'appartenance. De manière empirique, si plus de 20 % des observations sont mal classées au terme du processus, la typologie construite n'est probablement pas solide (nombre de classes inadapté ou absence de classifiabilité des observations). En outre, l'examen des « mal classés » permet d'iden-

tifier les observations périphériques à deux ou plusieurs classes ; compte tenu du profil de ces observations, leur migration vers une autre classe doit donc correspondre à une logique de proximité : les transferts s'effectuent entre classes dont les contenus sont logiquement voisins. L'absence d'une telle logique est le signe d'une qualité insuffisante de la typologie construite. Le tableau 5.6 présente les résultats observés sur la typologie d'un échantillon d'annonces publicitaires, où les fonctions discriminantes calculées permettent de bien classer 94 % des observations.

Tableau 5.6 – Réaffectation des observations à leurs classes d'origine (« table de classification »)

Groupes	n° 1	n° 2	n° 3	n° 4
Effectif des groupes initiaux	Effectif des groupes prédits avec les fonctions discriminantes :			
	24	21	29	26
n° 1 : 27	24 (88,9%)*	0	2 (7,4%)	1 (3,7%)
n° 2 : 22	0	20 (90,9%)	1 (4,6%)	1 (4,6%)
n° 3 : 27	0	1 (3,7%)	26 (96,3%)	0
n° 4 : 24	0	0	0	24 (100%)

*répartition en pourcentage de l'effectif initial entre les groupes reconstitués

Pour conclure, la valeur explicative d'une typologie et son opérationnalité pour le décideur reposent sur le soin apporté dans sa réalisation : choix et contrôle des variables utilisées, qualité d'analyse et dimension critique nécessairement portés sur les résultats obtenus. Processus itératif, la réalisation d'une typologie n'a rien d'automatique et implique de mettre en œuvre les moyens nécessaires en temps et en réflexion pour s'assurer de la validité opérationnelle des structures ainsi dégagées.

Pour aller plus loin – Calculer la distance entre groupes – exemple « ONG caritative »

Dans l'exemple « ONG caritative » la distance entre les profils des 47 secteurs a été calculée avec la Distance Euclidienne sous forme d'une matrice carrée et symétrique de 47 lignes et autant de colonnes. L'étape suivante consiste à réaliser la classification à l'aide de la méthode du « voisin le plus proche » (*nearest neighbour*). À ce stade chaque secteur est considéré comme un groupe. La construction de l'arbre commence par l'identification des deux premiers secteurs pour lesquels la distance euclidienne est minimale. Cette distance minimale (0,021) correspond au couple secteur 19 - secteur 15. Ce couple constitue alors la première classe (groupe) et donne la première branche du dendrogramme. La hauteur de cette branche sera égale à la distance euclidienne. À cette étape, les deux secteurs sont fusionnés en une seule classe et la matrice des distances aura donc une ligne et une colonne de moins. En utilisant la méthode du « plus proche voisin », les distances entre la classe déjà formée et chacun des 45 secteurs individuels qui restent seront égales à la valeur minimale des distances des secteurs 15 et 19 qui composent la nouvelle classe.

L'étape suivante consiste à rechercher le couple qui a la distance minimale. Deux cas de figure sont alors possibles : soit un nouveau secteur (différent de 15 et 19 qui ont ☞

© Dunod – Toute reproduction non autorisée est un délit.

déjà fusionnés pour donner la classe 15-19) vient rejoindre la nouvelle classe en donnant ainsi une classe de 3 secteurs ; soit deux secteurs créent une nouvelle classe. Dans le premier cas, la première branche est modifiée, alors que dans le deuxième, une nouvelle branche est créée intégrant les deux secteurs nouvellement fusionnés. Quel que soit le cas, la matrice des distances perd encore une ligne et une colonne. Cette procédure est répétée 46 fois (n-1). Le tableau 5.7 affiche les secteurs ou les classes combinés à chaque étape, les distances entre les observations ou les classes en cours de création, et le dernier niveau de classe auquel un secteur a rejoint la classe en appliquant la méthode d'agrégation du « voisin le plus proche ». La figure 5.3 compare les dendrogrammes en utilisant plusieurs méthodes d'agrégation.

Tableau 5.7 – Planning des agglomérations

Étape	Classe combinée		Coefficients	Étape de première apparition de la classe		Étape suivante
	Classe 1	Classe 2		Classe 1	Classe 2	
1	15	19	0,021	0	0	9
2	5	6	0,022	0	0	12
3	13	17	0,024	0	0	4
4	12	13	0,031	0	3	5
5	12	18	0,033	4	0	15
…	…	…	…	…	…	…
43	1	12	0,553	30	19	44
44	1	25	1,943	43	32	45
45	1	34	2,425	44	42	46
46	1	43	2,668	45	41	0

On peut remarquer qu'il existe des différences assez notables entre les dendrogrammes selon la méthode utilisée ; il est donc important d'identifier la structure la plus stable fixant l'appartenance des observations aux mêmes groupes indépendamment de la méthode d'agrégation.

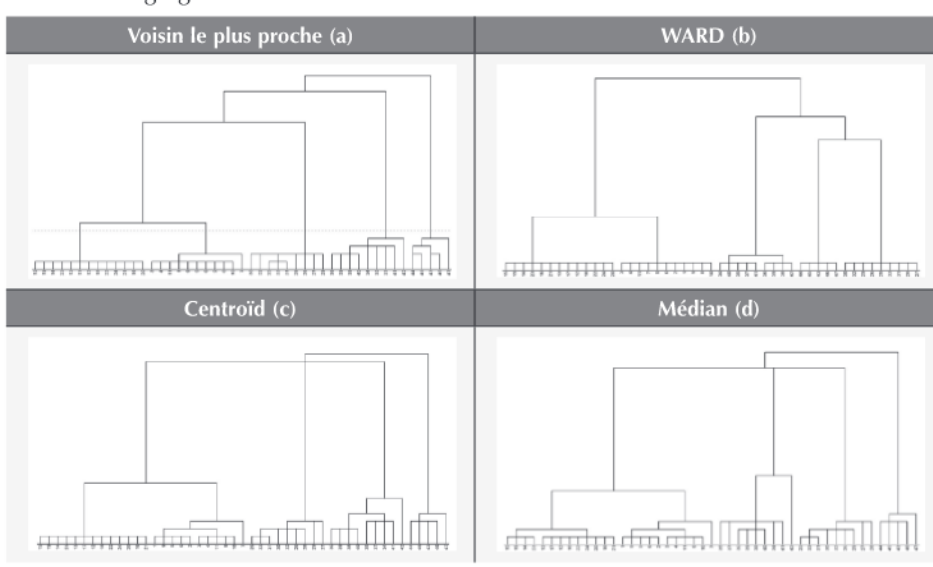


Figure 5.3 – Dendrogramme utilisant plusieurs méthodes d'agrégation

3 Illustration avec le cas « ONG »

3.1 Analyser la stabilité de la typologie

La distance euclidienne est l'indice de similarité choisi car l'examen des profils qualifiant les secteurs géographiques montre des disparités dans leur distribution. Pour éprouver la stabilité du modèle typologique à retenir, deux indices de proximité entre les groupes ont été utilisés : l'indice de Ward (le plus en adéquation avec la distance euclidienne – *cf.* section 2, § 1.1) et celui du « voisin le plus proche ».

Le tableau 5.8 présente les résultats concernant plusieurs structures typologiques entre lesquels choisir. Ces structures sont obtenues, d'une part avec une classification hiérarchique ascendante et, d'autre part – à titre d'illustration compte tenu de la faible taille de l'échantillon – avec une méthode non hiérarchique : celle des « centres mobiles ». Pour la méthode hiérarchique, on constate que le nombre de groupes des modèles diffère un peu selon l'indice de distance : de 4 à 6 avec le « voisin le plus proche » et de 3 à 5 avec « Ward ». La méthode des centres mobiles suggère trois modèles avec 2, 4 ou 5 groupes.

Tableau 5.8 – Comparaison des modèles typologiques selon la méthode de classification et les indices de fusion (fichier « ONG »)

	(a) Nombre de classes		(b) Effectif des classes		(c) Affectations identiques selon les indices (en %)
Indices de proximité entre groupes	Voisin le plus proche	Ward	Voisin le plus proche	Ward	
Méthode Hiérarchique (CAH)		3	-	24 ; 14 ; 9	-
	4	4	24 ; 9 ; 5 ; 9	24 ; 5 ; 9 ; 9	100 %
	5	5	11 ; 13 ; 9 ; 5 ; 9 ;	11 ; 13 ; 5 ; 9 ; 9 ;	100%
	6		11 ; 13 ; 9 ; 5 ; 4	-	-
Méthode non Hiérarchique (centres mobiles)	2	2	38 ; 9	38 ; 9	100 %
	4	4	24 ; 9 ; 5 ; 9	24 ; 5 ; 9 ; 9	100%
	5	5	13 ; 11 ; 9 ; 5 ; 9	11 ; 13 ; 5 ; 9 ; 9 ;	100%

Il apparaît que deux modèles typologiques – en 4 et en 5 classes – sont communs aux deux indices appliqués. Mais la question est de savoir si ces deux classes sont de taille équivalente quels que soit la méthode et l'indice de proximité utilisés et, dans l'affirmative, si elles contiennent à peu près les mêmes observations. La colonne (b) du tableau 5.8 permet de constater que les effectifs sont identiques et que les attributions sont rigoureusement les mêmes (colonne c), exprimant ainsi une stabilité maximale (mais c'est loin d'être aussi tranché dans de très nombreuses

© Dunod – Toute reproduction non autorisée est un délit.

études où un certain niveau de fluctuation des observations entre groupes est fréquent). Par ailleurs, la stabilité de la distance mesurée entre les classes dans les modèles comparés est aussi un indicateur de la stabilité de la typologie. On constate ici qu'elles sont stables (tableau 5.9).

Tableau 5.9 – Évolution de la distance entre les classes selon les modèles typologiques comparés

Distances entre types selon le modèle à 4 ou 5 classes								
4 classes	1	2	3	5 classes	1	2	3	4
2	2,948			2	0,860			
3	4,410	3,684		3	2,918	3,03		
4	4,705	4,449	5,568	4	4,751	4,141	3,684	
				5	4,833	4,630	4,449	5,568

Ces résultats signifient que les données qu'exploite cette ONG sont très fortement structurées et confirment aussi leur « classifiabilité ». Le choix du nombre de types va donc dépendre du degré de cohésion interne propre à chaque groupe et des caractéristiques qui les différencient de manière significative.

3.2 Analyser la validité des groupes et les décrire

La validité dépend à la fois de l'homogénéité des groupes et de la significativité des variables utilisées dans l'élaboration de la classification. L'homogénéité peut être appréciée par le niveau de variance à l'intérieur des groupes, aussi faible que possible (variance intragroupe) et, parallèlement, par la disparité entre les groupes laquelle s'apprécie par le ratio de la variance intergroupe/variance totale : plus il est proche de l'unité et plus il signifie que les groupes sont globalement très différenciés. Le tableau 5.10 montre, sur ces deux critères, que les deux modèles sont très satisfaisants : rapportée à la variance totale, on observe une très faible variance au sein des groupes.

Cependant, si le passage en 5 classes améliore mécaniquement l'homogénéité, ce n'est qu'en faible proportion (1,54 %), surtout comparé au passage de 2 à 4 classes qui améliore le degré d'homogénéité de 89 % (2,831/5,86). Cela conduit donc à renoncer au modèle en 2 classes, d'autant plus que la variance de la seule classe 1 y représente près de la moitié de la variance totale (48,8 %). Enfin, les distances entre les groupes sont bien marquées, à l'exception des deux premières classes du modèle en 5 classes : faut-il alors les différencier l'une de l'autre ?

Tableau 5.10 – Niveau des variances intra- et inter- groupes pour les modèles typologiques comparés

Classe	Variance intragroupe		Effectifs		Variance totale intergroupe	
	4 classes	5 classes	4 classes	5 classes	4 classes	5 classes
(a)	0,108	0,010	24	11		
(b)	-	0,004	-	13		
(c)	0,009	0,009	9	9		
(d)	0,014	0,014	5	5		
(e)	0,025	0,025	9	9		
Total	0,156	0,062			5,705	5,798
Ratio : Variance intergroupe/variance totale					0,973	0,099

3.3 Vérifier la valeur discriminante de la typologie

On peut encore éprouver la validité d'un modèle typologique avec l'analyse de variance pour vérifier le niveau de significativité des variables utilisées dans la partition. Et, pour finir, on peut s'assurer que le modèle permet de bien prévoir l'appartenance d'une observation à une classe sur la base d'une analyse discriminante.

– Le tableau 5.11 présente les résultats de l'analyse de variance sur les composantes factorielles à la base de la partition. Elles sont toutes « naturellement » significatives puisqu'elles ont généré la typologie (valeur du F et probabilité critique). Néanmoins, chaque facteur ne discrimine pas l'ensemble des types générés ; ce traitement permet de percevoir et de valider leur rôle spécifique dans la construction de la partition par comparaison des groupes non distingués.
– L'analyse des variables initiales (à la source des composantes factorielles) complète utilement cette approche globale. Sur les neufs variables utilisées pour diagnostiquer le « comportement » différentiel des secteurs, on identifie celles qui discriminent l'ensemble des groupes (en l'occurrence ici : 4 variables) et, quand ce n'est pas le cas, repérer s'il y a indifférenciation systématique entre groupes et avec quelles variables. En l'espèce, la « variation du nombre de donateurs » ne différencie guère les groupes 3 et 4 dans le modèle en 4 classes. Une lecture globale permet donc de repérer les variables sur lesquelles les groupes se distinguent significativement les uns des autres (tableau 5.11) et les variables qui caractérisent significativement les groupes en comparant leur moyenne à la moyenne générale de l'échantillon étudié (tableau 5.12).
– Un dernier test permet d'évaluer la qualité des structures typologiques : la prévisibilité de l'appartenance à une classe à partir de l'analyse du profil d'une observation quelconque. Une analyse discriminante permet de construire ce modèle prédictif et de vérifier quel est le taux de bonne attribution (cf. l'exemple du tableau 5.6 ou « table de classification » qui présente les taux d'attribution aux classes à partir des fonctions discriminantes calculées). Dans le cas ONG, quel que

© Dunod – Toute reproduction non autorisée est un délit.

soit le modèle retenu en 4 ou 5 classes, toutes les observations sont parfaitement bien classées (100 % de bonne attribution) : ce résultat valide les deux structures typologiques entre lesquelles l'analyste devra choisir.

Tableau 5.11 – Résultats des analyses de variances pour les deux modèles typologiques comparés

variable	Valeur du F et probabilité		Valeur du F et probabilité		Nb groupes discriminés	groupes non discriminés	Nb groupes discriminés	groupes non discriminés
Modèle de classification :	**4 cl**		**5 cl**		**4 cl**		**5 cl**	
Variation donateur 2013/2012	367,6	0,000	296,54	0,000	3	3-4	4	4-5
Variation donateur 2014/2013	711,41	0,000	616,33	0,000	4		5	
Variation donateur 2015/2014	110,07	0,000	855,16	0,000	3	3-4	5	
Variation don moyen 2013/2012	329,24	0,000	536,18	0,000	4		4	1-3
Variation don moyen 2014/2013	375,65	0,000	488,70	0,000	4		5	
Variation don moyen 2015/2014	349,99	0,000	327,86	0,000	4		5	
Variation montant 2013/2012	428,08	0,000	350,22	0,000	4		5	
Variation montant 2014/2013	298,02	0,000	246,62	0,000	4		5	
Variation montant 2015/2014	58,68	0,000	409,81	0,000	3	1-2	4	2-4
Composante 1	2440,74	0,000	2345,32	0,000	3	2-4	4	3-5
Composante 2	599,66	0,000	895,56	0,000	4		4	2-4
Composante 3	115,07	0,000	331,02	0,000	3	1-3	5	

Tableau 5.12 – Comparaison du profil des groupes sur les variables de « comportement »

% d'évolution entre années	Modèle à 4 classes				Modèle à 5 classes				
variable	g,1	g,2	g,3	g,4	g,1	g,2	g,3	g,4	g,5
Variation donateur 2013/2012	1,0	- 12,5	- 18,7	- 19,9	0,02	1,7	- 12,5	- 18,7	- 19,9
Variation donateur 2014/2013	1,2	- 8,6	6,0	- 12,5	0,07	1,6	- 8,6	6,0	- 12,5
Variation donateur 2015/2014	- 4,4	- 10,9	- 16,7	- 18,6	- 1,5	- 6,9	- 10,9	- 16,7	- 18,6

% d'évolution entre années	Modèle à 4 classes				Modèle à 5 classes				
variable	g,1	g,2	g,3	g,4	g,1	g,2	g,3	g,4	g,5
Variation don moyen 2013/2012	2,9	4,3	17,5	- 8,4	4,6	1,5	4,3	17,5	- 8,4
Variation don moyen 2014/2013	1,6	7,5	- 10,8	- 4,4	2,8	0,6	7,5	- 10,8	- 4,4
Variation don moyen 2015/2014	2,2	9,9	14,3	- 1,5	2,9	1,5	9,9	14,3	- 1,5
Variation montant 2013/2012	3,9	- 8,8	- 4,4	- 26,8	4,8	3,3	- 8,8	- 4,4	- 26,8
Variation montant 2014/2013	2,8	- 1,8	- 5,4	- 16,3	3,5	2,3	- 1,8	- 5,4	- 16,3
Variation montant 2015/2014	- 2,3	- 2,1	- 4,8	- 19,9	1,4	- 5,5	- 2,1	- 4,8	- 19,9

■ Les variables initiales sur lesquelles les groupes concernés ne se différencient pas.

3.4 Choix du modèle final et prise de décision

Bien que les deux modèles typologiques proposés soient robustes, il reste à en choisir un pour poser le diagnostic le plus pertinent et recommander en conséquence les actions les mieux adaptées aux « comportements » des secteurs concernés. En fait, dans le cas présent, en passant de 4 à 5 groupes, la différence ne concerne qu'une classe : la classe 1 du modèle à 4 classes se scinde en deux groupes (classes 1 et 2 du modèle en 5 classes) puisque toutes les autres classes restent identiques.

Si le modèle en 4 classes est assez déséquilibré (une seule classe représente la moitié des observations), il faut se demander si, d'un point de vue opérationnel, sa décomposition en deux classes (d'effectif presque identique : respectivement 27,6 % et 23,4 %) permet de mieux affiner le diagnostic et d'adapter les recommandations. Pour le savoir, il faut donc analyser les contenus spécifiques à chaque groupe, dont l'étude comparative entre les deux modèles va permettre de décider lequel présente le plus grand intérêt tactique.

Le tableau 5.13 synthétise les écarts entre les groupes. Il apparaît que les groupes 1 et 2 du modèle en 5 classes représentent plus des 2/3 des donateurs tous secteurs confondus. Si le « comportement » de ces deux groupes est assez voisin, on relève que les secteurs qu'ils englobent présentent sur la majorité des critères une évolution globalement favorable au cours des années, sauf en 2015. Il faut donc stabiliser la situation et faire en sorte que la tendance, plutôt favorable antérieurement, perdure.

Cela peut conduire à maintenir les deux groupes fusionnés et privilégier le modèle en 4 classes. Cependant, en 2015, le nombre de donateurs baisse sensible-

© Dunod – Toute reproduction non autorisée est un délit.

ment pour tous les groupes (et fortement pour les groupes 4 et 5). Or, les différences entre les groupes 1 et 2 concernent précisément cette année 2015, où l'augmentation du don moyen du groupe 1 compense la perte de donateurs pour maintenir un montant global stable, alors que la stabilité du don moyen au cours des ans dans le groupe 2 ne compense pas suffisamment la perte de donateurs, entraînant une baisse du montant global. En plus des actions à envisager pour fidéliser les donateurs actuels (groupes 1 et 2), il faudrait que le don moyen du groupe 2 puisse augmenter.

Sur les trois autres types, on constate un comportement assez voisin caractérisé par une baisse conséquente du nombre de donateurs, avec une augmentation du don moyen qui ne suffit pas pour compenser la perte de donateurs (sauf pour le groupe 3). Ces trois types se distinguent en termes de degré. Il faut donc chercher à comprendre sur le terrain les raisons de cette régression pour envisager les solutions adaptées. C'est en fait sur la conquête de nouveaux donateurs qu'il faut travailler (groupes 3 et 4), ou sur la fidélisation au don (groupes 4 et 5) en cherchant à récupérer ceux qui ne donnent plus (après en avoir étudié les raisons).

Tableau 5.13 – Synthèse de l'évolution des variables selon les groupes

% d'évolution entre années variables	Modèle à 4 classes				Modèle à 5 classes				
	g.1	g.2	g.3	g.4	g.1	g.2	g.3	g.4	g.5
Variation donateur 2013/2012	=	-	- -	- -	=	=	-	- -	- -
Variation donateur 2014/2013	=	-	+	-	=	=	-	+	-
Variation donateur 2015/2014	-	-	--	--	=	-	-	--	--
Variation don moyen 2013/2012	=	+	++	-	+	=	+	++	-
Variation don moyen 2014/2013	=	+	-	-	+	=	+	-	-
Variation don moyen 2015/2014	=	+	+	=	+	=	+	+	=
Variation montant 2013/2012	=	-	-	---	+	+	-	-	---
Variation montant 2014/2013	=	=	-	--	+	+	=	-	--
Variation montant 2015/2014	=	=	-	--	=	-	=	-	--

Pour aller plus loin – Présentation de la logique de l'analyse discriminante et des fonctions discriminantes

L'analyse discriminante est une technique statistique qui fait partie des analyses factorielles : son objectif est d'affecter les observations (individus, objets, etc.) à des groupes connus par avance en utilisant les variables quantitatives. La technique de l'analyse discriminante linéaire a été introduite en 1936 par Ronald Fisher et son principe consiste à trouver des combinaisons linéaires de variables quantitatives pour permettre l'affectation des observations (individus, objets, …) dans chacun des groupes constitués.

L'analyse discriminante peut être une technique descriptive (calcul de facteurs à partir des variables prédictives permettant de visualiser graphiquement les proximités entre les observations sur le(s) plan(s) factoriel(s)), ou prédictives (permettant de prédire à quel groupe appartient une observation en fonction des variables prédictives). L'analyse discriminante permet donc de vérifier quel est le degré de bonne attribution des observations aux groupes déjà constitués à l'aide des méthodes typologiques. Bien que sa mise en œuvre soit assez complexe, l'analyse factorielle discriminante n'est rien d'autre qu'une ACP du nuage des *k* centres de gravité (k étant le nombre de catégories d'une variable qualitative désignant les groupes, c'est-à-dire le nombre des groupes d'affectation).

L'affectation des observations à des groupes définis *a priori* s'effectue à l'aide des « fonctions discriminantes » qui sont chacune une combinaison linéaire des variables prédictives. Elle prend alors le nom de l'analyse discriminante linéaire.

La fonction linéaire discriminante pour *q* variables prédictives s'écrit comme une fonction de régression :

$u = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q$

avec u comme variable à prédire ; X_1, X_2, …X_q – comme variables prédictives et β_0, β_1, β_2, …β_q étant les coefficients à estimer d'après les valeurs observées sur l'échantillon. L'analyse factorielle discriminante (AFD) cherche à trouver les combinaisons linéaires qui représentent des groupes homogènes (variation minimale dans les groupes) et bien distincts les uns des autres (variation maximale entre les groupes).

Le nombre de facteurs extraits par l'AFD est égal à min{q, k –1} et chaque axe factoriel a un pouvoir discriminant (λ [0;1]) qui varie entre 0 et 1 : 0 si la projection les centres de gravité des groupes projetés sur un facteur sont confondus et 1 (cas idéal) si la dispersion dans les groupes est nulle et donc si les centres de gravité des groupes projetés sur un facteur sont bien distincts). Les valeurs propres des facteurs μ [0;+∞[sont définies par les valeurs propres λ avec la relation $\mu = \lambda/(1-\lambda)$.

Pour chaque fonction linéaire discriminante, la valeur propre représente le rapport de la somme des carrés des écarts *entre* les groupes sur la somme des carrés des écarts *dans* les groupes.

© Dunod – Toute reproduction non autorisée est un délit.

L'ESSENTIEL

■ ***Sélectionner les critères pertinents pour construire le profil de comparaison***

Ne pas prendre toutes les variables, mais choisir seulement celles qui ont du sens en regard des objectifs et de la finalité de la typologie à réaliser. S'assurer que certains traits ne sont pas involontairement surreprésentés par rapport à d'autres et que tous les aspects pertinents et importants sont bien présents.

■ ***Rendre les variables comparables et de poids identique***

Vérifier que les variables introduites sont indépendantes et que les métriques utilisées ne donnent pas artificiellement plus de poids à certaines d'entre elles. L'usage de l'analyse factorielle métrique (ACP) ou non métrique (ACM) est une pratique recommandée pour donner à toutes les variables un statut équivalent en utilisant les scores factoriels.

■ ***Choisir l'indice de comparaison entre les observations adapté aux variables***

En premier lieu, le choix de l'indice de similarité (ou de distance) entre les profils est dicté par la nature métrique ou non des variables et, en second lieu, par le type de mesure de comparaison entre les observations que l'analyste veut privilégier pour construire sa typologie.

■ ***Appliquer un protocole de classification pour optimiser la typologie construite***

L'association d'une méthode hiérarchique pour déterminer le nombre de classes (à partir du « dendrogramme ») et d'une méthode non hiérarchique pour optimiser la répartition des observations entre les groupes est recommandable, en utilisant l'indice de Ward comme mesure de la distance entre les groupes. L'analyse d'un certain nombre d'indicateurs (ratios de variance intra- et inter- groupes et distances entre groupes) permet d'orienter le choix vers une partition.

■ ***Valider la pertinence et l'opérationnalité de la typologie retenue***

L'analyse statistique du rapport des variables actives et illustratives avec la typologie dans sa globalité, puis avec chaque groupe permet de confirmer la classifiabilité des observations et la signification spécifique des groupes les uns par rapport aux autres. Enfin, l'usage des fonctions discriminantes permet de vérifier la stabilité de la typologie et de prédire l'affectation d'une observation quelconque à l'un des groupes constitués.

Chapitre 6

Comment réaliser et valider des prévisions

OBJECTIFS

- Savoir situer la spécificité d'un type de modèle régressif, linéaire ou non, dans une démarche prévisionnelle pour formuler avec pertinence les objectifs de l'étude statistique à conduire.
- Déterminer les conditions d'application des modèles de régression linéaire simple et multiple et être capable d'évaluer la qualité de la prévision.
- Comprendre la logique des méthodes prévisionnelles fondées sur l'analyse des séries temporelles et pouvoir identifier la tendance générale (ou *trend*) de la série et la présence d'une éventuelle composante saisonnière.

SOMMAIRE

SECTION 1 La régression linéaire simple (RLS)

SECTION 2 La régression linéaire multiple (RLM)

SECTION 3 La prévision fondée sur des séries temporelles

Si l'intensité de la liaison entre deux variables métriques X et Y est mesurée par le coefficient de corrélation (où les deux variables ont un rôle symétrique), il ne rend pas compte d'une relation de cause à effet. Pourtant, il est fréquent qu'un décideur souhaite prévoir le résultat d'une variable métrique *Y* (appelée *variable à expliquer* ou *variable dépendante*) à partir des valeurs connues d'une ou plusieurs autres variables X_p (appelées *variables explicatives, indépendantes* ou encore *prédicteurs*). L'objectif de prévoir les valeurs d'une variable à partir d'un ensemble d'autres variables correspond à une problématique statistique : *l'analyse de régression*.

Si la relation entre les deux variables X et Y est parfaite – coefficient de corrélation linéaire égal à 1 ou -1 – alors il est possible connaître avec précision les valeurs de Y pour chacune des valeurs de X. Sur le plan statistique, une telle dépendance implique que les observations, représentées sous la forme d'un nuage de points, se trouvent toutes alignées sur une droite. Par exemple, c'est le cas de la prévision de la masse d'un cylindre d'un diamètre donné à partir sa longueur. Mais, en sciences de gestion, un tel cas idéal ne se rencontre pratiquement jamais. Le plus souvent, il existe un certain écart entre la valeur prédite par le modèle théorique (une droite de régression, par exemple) et la valeur observée. Cet écart est dû à de très nombreux effets aléatoires qui conditionnent le phénomène étudié et qui sont impossibles à mesurer.

Il existe une diversité de techniques de régression dont le choix dépend de la nature métrique ou non des variables, du nombre des variables prédictives et de la forme, linéaire ou non, de leur relation. Mais la différenciation entre « variable à expliquer » et « variable explicative » est à la base d'une logique prédictive qui – contrairement à la corrélation – formalise le lien entre ces variables et permet ainsi au décideur, ou au manager, de répondre à ses besoins en matière de prévision, ou d'effectuer nombre de calculs économiques. L'objectif de ce chapitre est précisément de présenter cette logique prévisionnelle, en s'appuyant d'abord sur la régression simple (section 1), avant de développer une pratique de la prévision fondée sur la régression multiple (section 2) et l'analyse des séries temporelles (section 3).

Section 1 LA REGRESSION LINÉAIRE SIMPLE (RLS)

L'exemple suivant permet d'illustrer le principe et l'usage de la régression linéaire simple.

Cas d'illustration

Le consultant indépendant BODOT

Bodot est une entreprise conseil dont l'un des pôles est spécialisé dans l'analyse des activités de la distribution. À ce titre, il est consulté par d'importants clients souhaitant avoir un avis sur la productivité de leurs implantations commerciales dont l'effectif en personnel est conséquent. Il lui faut donc se doter d'un modèle de prévision afin d'estimer le CA théorique susceptible de lui servir de référence en fonction du personnel employé (car on lui pose la question en terme de productivité).

À cette fin, il s'appuie sur l'enquête *Points de Vente* réalisée en 2009 par l'INSEE qui a permis de mesurer, entre autres choses, le chiffre d'affaires des grandes surfaces (tous secteurs confondus) et le nombre de personnes occupées[1]. Le tableau ci-dessous présente, pour la France, un extrait des résultats relatifs aux secteurs des grandes surfaces ayant entre 4 000 et 30 000 « personnes occupées » (nombre exprimé en équivalent temps plein).

Tableau 6.1 – Évolution du CA en fonction de l'effectif des grandes surfaces

Personnes occupées, en milliers	4	9	19	20	26	30
Chiffre d'affaires, en millions d'euros	1 280	1 920	4 560	2 940	5 120	6 220

Deux de ses clients lui demandent si le CA qu'ils réalisent avec leurs magasins est comparable à celui des magasins employant un nombre de « personnes occupées » à peu près équivalent. Le premier (INDEP) est un indépendant, gestionnaire d'un ensemble de magasins de détail, « occupant » 12 000 personnes et réalisant un CA annuel de 2 800 millions d'euros. Le second (ENSE) est une enseigne qui emploie 23 000 personnes dans ses grandes surfaces et réalise un CA de 3 100 millions d'euros. Bodot va tenter de répondre précisément à leur souci avec son modèle de prévision.

1 Principe, données et conditions d'application de la RLS

La régression linéaire simple (RLS) est le cas le plus élémentaire qui traduit la relation linéaire d'une variable dépendante (Y) par rapport à une autre variable indépendante (X). S'il existait une relation parfaite entre les deux variables, tous les points seraient parfaitement alignés sur une droite, appelée *droite de régression* et, dans ce cas précis, il s'agirait d'un modèle déterministe. En gestion, même s'il existe une dépendance linéaire dans le plan formé par les deux axes X et Y, les points se

1. Selon l'INSEE, les « personnes occupées » regroupent les effectifs salariés et non-salariés, ainsi que le personnel mis à disposition du point de vente et le personnel intérimaire, mais en excluant le personnel du point de vente mis à disposition ailleurs.

© Dunod – Toute reproduction non autorisée est un délit.

dispersent plutôt à l'intérieur d'un « nuage de points » plus ou moins étiré. L'objectif de ce modèle statistique est de trouver la droite qui décrirait au mieux ce nuage. Il s'agit donc d'un modèle probabiliste. Lorsque la dépendance de Y par rapport à X est linéaire pour une population donnée, le modèle (*i.e.* l'équation de la droite) peut être décrit par l'équation suivante :

$$Y = \beta_0 + \beta_1 X$$

où X est la variable indépendante (ou variable exogène, prédicteur) ; Y est la variable dépendante (ou variable endogène, variable de réponse) ; β_0 est l'ordonnée à l'origine qui exprime la valeur de Y quand $X = 0$ et β_1 est la pente qui représente la variation de la valeur de Y lorsque X varie d'une unité. La valeur observée Y_i de la variable dépendante peut être alors exprimée par la valeur X_i à l'aide de la relation :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

où ε_i représente l'écart entre la valeur observée et celle prédite par le modèle linéaire. La structure des données de la régression linéaire simple est présentée par le tableau 6.2. Il y a donc n valeurs de la variable indépendante X et autant de valeurs correspondantes de la variable dépendante Y.

Tableau 6.2 – Format des données de la régression linéaire simple

X	x_1	x_2	…	x_i	…	X_n
Y	y_1	y_2	…	y_i	…	Y_n

Remarque

À ce niveau, il est très important de noter que la variable indépendante X, contrairement à la variable dépendante, n'est pas une variable aléatoire parce que ses valeurs ne sont pas obtenues par un processus aléatoire mais sont fixées par l'analyste. D'ailleurs, le choix des valeurs de X a un effet important sur la qualité et la précision du modèle linéaire. Dans l'exemple INSEE, les valeurs de X n'explorent que les grandes surfaces ayant eu entre 4 000 et 30 000 personnes occupées. La valeur x = 0 signifie « aucune personne occupée », ce qui n'a pas vraiment de sens.

Le caractère aléatoire de la variable indépendante Y est dû aux erreurs. Les valeurs Y_i de la variable dépendante représentent les moyennes (espérances) des distributions conditionnelles $E(Y_i)$. Pour chacune des valeurs fixes de X (*e.g.* nombre des personnes occupées), il existe une multitude de valeurs de Y (*e.g.* chiffres d'affaires des grandes surfaces) en raison de plusieurs facteurs qui ne sont pas contrôlables *a priori*. Cette multitude de valeurs de Y correspond donc à la distribution de la probabilité de la variable Y pour chacune des valeurs de la variable X. Il s'agit alors de distributions conditionnelles de probabilités : $p(Y_i | X_i)$. En réalité, ces distributions de probabilités de Y sont toutes plus ou moins différentes et elles sont caractérisées par des espérances (moyennes) et des variances différentes.

Focus

Conditions d'application de la régression linéaire

- Le prédicteur X doit avoir une corrélation statistiquement significative avec la variable prédite Y,
- Les variables aléatoires ε_i doivent être distribuées normalement, ce qui implique automatiquement la normalité des variables aléatoires Y_i,
- Les variables aléatoires ε_i doivent avoir toutes la même variance, notée σ^2 (cette condition est appelée homoscédasticité des variances). Cette contrainte s'exprime : $Var(\varepsilon_i) = Var(Y_i) = \sigma^2$
- Les variables aléatoires ε_i doivent toutes avoir une espérance mathématique nulle. Ces trois dernières conditions signifient que les résidus ε_i doivent suivre une loi normale de moyenne nulle et de variance souvent inconnue σ^2 : $\varepsilon_i \rightarrow N(0; \sigma^2)$.
- La moyenne μ_i de chaque distribution $p(Y_i | X_i)$ doit se situer sur la droite de régression.
- Les variables aléatoires ε_i sont indépendantes entre elles. Cette condition s'exprime comme : $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$. L'indépendance des ε_i implique l'indépendance des Y_i parce que $Cov(Y_i, Y_j) = Cov(\beta_0 + \beta_1 x_i + \varepsilon_i, \beta_0 + \beta_1 x_j + \varepsilon_j) = Cov(\varepsilon_i, \varepsilon_j)$

La figure 6.1 schématise la forme des distributions conditionnelles des Y_i selon les suppositions faites dans le cas de la régression linéaire simple.

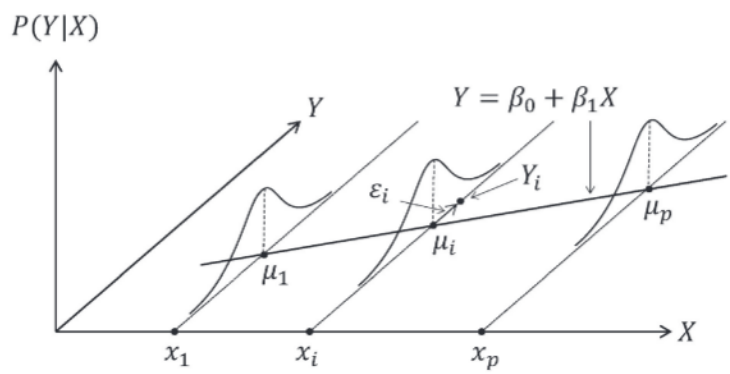


Figure 6.1 – Distribution de Y selon les hypothèses de la régression simple

La population étudiée étant rarement connue, il n'est pas possible de déterminer directement les coefficients β_0 et β_1 de la droite de régression. Il faut alors recourir à un échantillon de taille *n*, représentatif de la population étudiée, qui se présente sous la forme de paires d'observations : $(x_1, y_1), (x_2,y_2),\ldots,(x_i,y_i),\ldots,(x_n,y_n)$.

© Dunod – Toute reproduction non autorisée est un délit.

Dans l'exemple introductif, pour chacune de 6 valeurs prises par le prédicteur *X* (personnes occupées), il y a un grand nombre de magasins dont le chiffre d'affaires diffère plus ou moins de l'un à l'autre. À partir des données détaillées pour chacun d'eux, il est alors possible de calculer la distribution du chiffre d'affaires pour chaque taille d'entreprise. La figure 6.2 trace le nuage de points entre les variables *Effectif* et *Chiffre d'affaires*. La dispersion des points indique que le nuage qui les représente peut être modélisé par une droite.

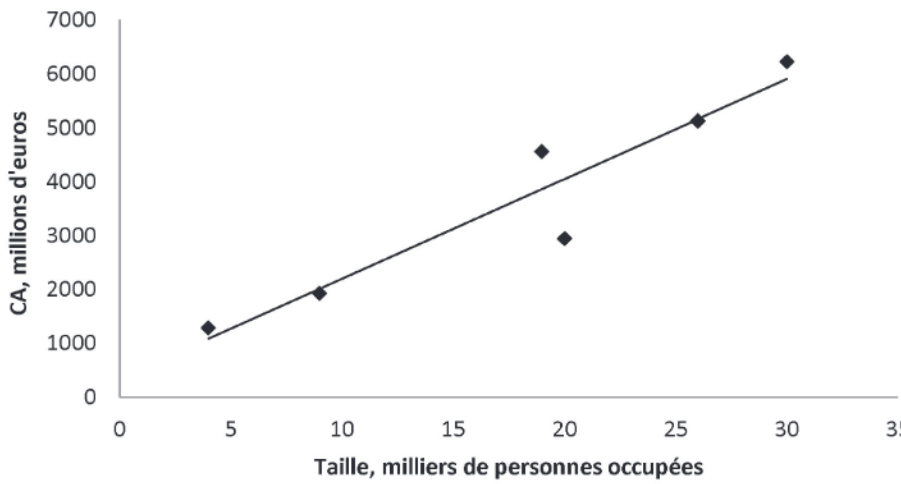


Figure 6.2 – Le nuage de points entre les variables *Effectif* et *CA*

Remarque

Quelle que soit la disposition des points dans le nuage résultant de l'association entre X et Y, il est toujours possible de calculer une droite de régression pour la fonction Y=f(X) et X=f(Y), mais pour interpréter les résultats, il faut considérer le sens logique de la relation causale associant les deux variables. Dans l'exemple traité, c'est bien l'effectif du magasin qui prédit son chiffre d'affaires, non l'inverse.

Pour passer de la formulation de l'hypothèse de dépendance linéaire de Y en X à la prise de décision managériale finale, il faut franchir plusieurs étapes dont les plus importantes sont énumérées ci-dessous :

- estimer les coefficients de régression à partir des données observées ;
- mesurer le pouvoir explicatif du modèle linéaire ;
- mesurer la signification globale du modèle linéaire ;
- mesurer la signification des coefficients de régression ;
- prendre la décision finale.

2 Estimation des coefficients de régression à partir des données observées

La première étape de la construction du modèle consiste à déterminer la valeur des coefficients de régression b_0 et b_1. Il s'agit de rechercher, empiriquement, la valeur propre à chacun de ces deux paramètres pour que les données calculées (estimées) par le modèle correspondent le mieux possible avec les données observées (le moins d'écart possible). C'est cette démarche qui permet d'appliquer aux données d'une population un modèle fabriqué sur les seules données d'un échantillon. Si la procédure a bien un caractère pragmatique, elle repose néanmoins sur une logique rigoureuse qui est développée ci-dessous.

Focus

Estimation des coefficients de régression

Soit un échantillon composé de n paires d'observations. Les n points dans l'espace cartésien formé par les axes X et Y peuvent être représentés graphiquement sous la forme d'un « nuage de points ». Si la répartition de cet ensemble de points tend à ressembler à une droite, alors l'équation de cette droite $\hat{y}$ capable de passer au plus près des points du nuage s'écrit :

$$\hat{y}_i = b_0 + b_1 x_i \ (1 \leq i \leq n)$$

où : x_i est la valeur i du prédicteur X ; $\hat{y}_i$ est valeur i de la variable prédite Y correspondant à une valeur de x_i, et se trouve sur la droite de régression ; b_0 est la valeur estimée de l'ordonnée à l'origine de la droite de régression (i.e. la valeur de Y quand X = 0) ; b_1 est valeur estimée de la pente de la droite de régression : elle mesure le niveau d'évolution de la valeur de Y quand la valeur de X augmente d'une unité.

Dans la réalité, les valeurs observées y_i ne se trouvent jamais sur la droite mais sont distribuées de manière aléatoire autour de la droite de régression, en raison de l'erreur d'échantillonnage mais aussi à cause d'un grand nombre de micro-facteurs qui ne peuvent être déterminés. Les valeurs y_i sont alors calculées par l'équation :

$$y_i = b_0 + b_1 x_i + e_i$$

où e_i est le terme qui permet de prendre en compte les erreurs ou *résidus*. Ces résidus représentent l'écart entre la valeur observée y_i et la valeur ajustée par la droite $\hat{y}_i$.

L'équation précédente peut être écrite sous la forme :

$$y_i = \hat{y}_i + e_i \Rightarrow e_i = y_i - \hat{y}_i$$

Il en résulte que, si toutes les erreurs e_i sont nulles, les points observés se trouvent alignés sur la droite de régression. Si les résidus e_i ne sont pas nuls, alors il faut déterminer l'équation de la droite pour laquelle ces erreurs sont les plus petites possibles à l'aide de la méthode des moindres carrés, laquelle consiste à déterminer les coefficients de régression b_1 et b_0 de manière à ce que la somme des carrés des écarts des résidus, notée SCE_R, soit la plus petite possible. Il est démontré que les coefficients de régression estimés par la méthode des moindres carrés sont déterminés par les formules suivantes :

$$b_1 = \frac{s_{xy}}{s_X^2} \text{ et } b_0 = \overline{y} - b_1 \overline{x}$$

© Dunod – Toute reproduction non autorisée est un délit.

Remarque

Le coefficient b_1 est une variable aléatoire car il est la combinaison linéaire de variables aléatoires. Comme les variables aléatoires Y_i sont, par hypothèse, distribuées normalement alors b_1 est également distribuée normalement. La droite de régression passe par un point particulier G de coordonnées $(\bar{x},\bar{y})$ appelé centre d'inertie du nuage des points. Les coordonnées $\bar{x}$ et $\bar{y}$ vérifient donc exactement le modèle (il n'y a pas de résidus).

Illustration du calcul des coefficients de corrélation avec l'exemple du cas Bodot. Le tableau 6.3 résume les sommes qui permettent de calculer facilement les statistiques descriptives des variables X et Y.

Tableau 6.3 – Statistiques descriptives des variables Effectif et CA

x_i	y_i	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$\hat{y}$	$(y_i-\bar{y})^2$	$(\hat{y}-\bar{y})^2$
4	1 280	33 507	196	5 728 044	1 078	40 842	6 736 249
9	1 920	15 780	81	3 074 178	2 005	7 198	2 783 858
19	4 560	887	1	786 178	3 859	491 792	34 369
20	2 940	-1 467	4	537 778	4 044	1 219 056	137 474
26	5 120	11 573	64	2 092 844	5 156	1 328	2 199 592
30	6 220	30 560	144	6 485 511	5 898	103 693	4 949 081
108	***3 673***	***90 840***	***490***	***18 704 533***		***1 863 910***	***16 840 624***
				SCE_T		SCE_R	SCE_M

On calcule d'abord les moyennes des variables X (Effectif) et Y (CA) : $\bar{x} = 108/6 = 18$ et $\bar{y} = 22040/6 = 3673$. On calcule ensuite les variances et les écarts-types observés des variables X et Y :

$$s_x^2 = \frac{490}{6} = 81,66$$

$$s_x = \sqrt{s_x^2} = \sqrt{81,667} = 97$$

$$s_y^2 = 18704533/6 = 3117422$$

$$s_y = \sqrt{3117422} = 1765,622$$

La covariance entre les variables X et Y est : $s_{xy} = \frac{90840}{6} = 15140$

Le coefficient de régression b_1 est : $b_1 = \frac{s_{xy}}{s_x^2} = \frac{15140}{81,667} = 185,388$

Alors que l'ordonné à l'origine b_0 est :

$b_0 = \bar{y} - b_1\bar{x} = 3673 - 185,388x18 = 336,354$

La fonction de régression est donc la suivante : $Y_i = 336,354 + 185,388X_i$. C'est la base du modèle prévisionnel dont Bodot souhaite de doter.

Le chiffre d'affaires des grandes surfaces peut être alors prédit par l'effectif selon le modèle linéaire $CA = 336,354 + 185,388 * Effectif$

Dans l'exemple, selon ce modèle théorique, le chiffre d'affaires des grandes surfaces ayant 19 000 personnes occupées serait de 336,354 + 185,388 * 19 = 3 859 millions d'euros ce qui représente un écart de 4 560 – 3 859 = 701 millions d'euros par rapport à la valeur réelle observée.

Remarque

Ces valeurs peuvent différer de celles obtenues par un logiciel qui divise les SCE par n-1 (pour un échantillon) et non *n* (pour la population).

3 Mesure du pouvoir explicatif du modèle linéaire

Une fois les coefficients de régression calculés, l'étape suivante consiste à vérifier si le modèle statistique établi est capable de reconstruire les données initiales observées, c'est-à-dire si la relation linéaire entre variable explicative et variable expliquée est bien avérée. Un indicateur synthétique permet de le vérifier : c'est le coefficient de détermination R^2.

La méthode des moindres carrés permet de calculer les coefficients b_0 et b_1 pour n'importe quel nuage de points. Il faut alors se demander si la qualité de l'ajustement est satisfaisante, c'est-à-dire si la droite de régression calculée rend compte avec fidélité des valeurs observées et les représente bien telles qu'elles étaient situées dans le nuage de points initial. Il s'agit donc de vérifier la capacité du prédicteur X à rendre compte de la plus grande partie des variations de la variable prédite Y.

Remarque

Ce questionnement est souvent exprimé ainsi : « dans quelle mesure les variations de X « expliquent » les variations de Y ? ». Dans le cadre d'un modèle régressif, le verbe « expliquer » n'exprime pas un lien de causalité entre X et Y, mais plus modestement : la *capacité à rendre compte* d'une variation.

© Dunod – Toute reproduction non autorisée est un délit.

Focus

Décomposition de la variance

Comme le montre la figure 6.3, l'écart entre la valeur observée et la moyenne d'Y ($y_i - \bar{y}$) qui représente l'écart total peut être exprimé comme la somme : d'une part, de l'écart de la valeur estimée par le modèle de régression $\hat{y}_i$ à la moyenne d'Y ($\hat{y}_i - \bar{y}$) et, d'autre part, de l'écart de la valeur observée y_i à l'estimation $\hat{y}_i$ ($y_i - \hat{y}_i$). On peut alors écrire :

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Il est démontré que la somme des carrés du terme de gauche est toujours égale à la somme des sommes des carrés des deux termes de droite.

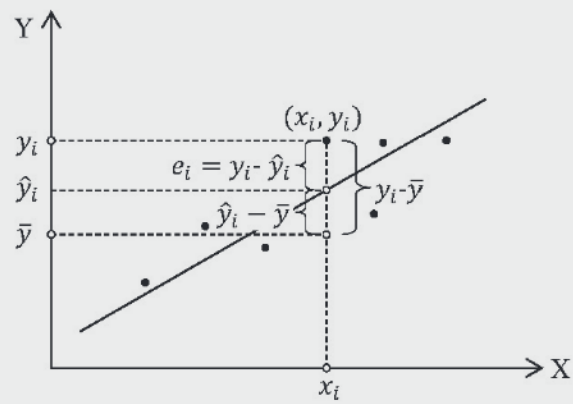


Figure 6.3 – Décomposition des écarts

La formule (1) donne la somme des carrés des écarts totaux de tous les points :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \qquad (1)$$

Le terme de gauche représente la Somme des Carrés des Écarts Totaux, notée SCE_T et traduit la variabilité totale de la variable dépendante, soit :

$$SCE_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Le terme du milieu représente la Somme des Carrés des Écarts expliqués par le modèle de régression, notée SCE_M : elle correspond à la part de la variabilité de la variable prédite dont le modèle de régression rend compte (« explique »), soit :

$$SCE_M = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

En remplaçant b_0 par $\bar{y} - b_1\bar{x}$ dans l'équation $\hat{y}_i = b_0 + b_1x_i$, on obtient $\hat{y}_i = \bar{y} - b_1\bar{x} + b_1x_i$, ce qui permet d'écrire (2) :

$$SCE_M = \sum_{i=1}^{n}[(\bar{y} - b_1\bar{x} + b_1x_i) - \bar{y}]^2 = b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (2)$$

☞

☞

Le terme de droite de l'équation (1) exprime la Somme des Carrés des Résidus (SCE_R) : elle représente la part de la variabilité de la variable dépendante qui n'est pas « expliquée » par le modèle de régression. En replaçant $\hat{y}_i$ par $b_0 + b_1 x_i$ nous obtenons cette nouvelle équation :

$$SCE_R = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^{n} e_i^2$$

La double origine des écarts possibles entre valeurs observées et prédites ayant été identifiée, la formule suivante, qui décompose les sources d'écarts, décrit la relation qui exprime la variabilité totale de la variable Y autour de sa moyenne :

$$SCE_T = SCE_M + SCE_R$$

De l'équation (2), on déduit que SCE_M est proportionnelle au carré de la pente de la droite de régression et à la variabilité du prédicteur. En conséquence, plus SCE_M sera grand et plus SCE_R sera petit, alors plus la droite de régression représentera fidèlement le nuage des points. Le rapport SCE_M/SCE_T est alors une bonne mesure de la proportion de la variabilité totale expliquée par la droite de régression. Ce rapport, noté R^2, est appelé coefficient de détermination et s'écrit :

$$R^2 = \frac{SCE_M}{SCE_T} = \frac{SCE_T - SCE_R}{SCE_T} = 1 - \frac{SCE_R}{SCE_T}$$

Dans le cas de la régression linéaire simple, il est démontré que le coefficient de détermination R^2 est égal au carré du coefficient de corrélation linéaire de Pearson :

$$R^2 = r^2$$

Le coefficient de détermination R^2 est un indicateur qui permet d'évaluer la qualité de la régression linéaire : il mesure le degré d'adéquation entre les données prédites (calculées) avec le modèle de régression et les données réellement observées.

Le coefficient de détermination R^2 varie entre 0 et 1. Plus il est proche de 1 et plus la droite de régression est fidèle aux points d'origine. Cependant l'interprétation de ce coefficient doit se faire avec prudence, car la qualité du modèle de régression dépend de l'ensemble des données analysées. En effet, le coefficient de détermination R^2 a l'inconvénient de croître avec le nombre des prédicteurs introduits dans le modèle régressif (voir section sur la régression linéaire multiple). C'est pour cette raison qu'il est nécessaire de calculer le *coefficient de détermination ajusté* qui tient compte du nombre de variables indépendantes comprises dans le modèle (une seule dans l'analyse de régression simple) :

$$R^2_{ajusté} = R^2 - \frac{p(1-R^2)}{n-p-1} = 1 - \frac{\left(1-R^2\right)(n-1)}{n-p-1}$$

où p est nombre des prédicteurs et n est la taille de l'échantillon.

© Dunod – Toute reproduction non autorisée est un délit.

Les statistiques descriptives appliquées au cas BODOT sont calculées dans le tableau 6.3. Les sommes des carrés des écarts sont : SCE_T = 18 704 533, SCE_R = 1 863 910 et SCE_M = 16 840 624. Les coefficients de détermination R^2 et $R^2_{ajusté}$ valent respectivement :

$$R^2 = 1 - \frac{SCE_R}{SCE_T} = 1 - \frac{1863910}{18704533} = 0,900$$

et

$$R^2_{ajusté} = 1 - \frac{\left(1 - R^2\right)(n-1)}{n-k-1} = 1 - \frac{(1-0,900)(6-1)}{6-1-1} \simeq 0,875$$

On peut alors conclure que le modèle linéaire construit rend effectivement compte de (ou « explique ») 88 % de la variabilité totale, soit un excellent résultat dans le cadre de cette étude.

Mais la significativité du R^2 dépend (entre autres choses) du nombre d'observations utilisées dans la construction du modèle ; or il est plutôt faible dans l'exemple des grandes surfaces : la question se pose alors de savoir s'il est statistiquement significatif.

Remarque

En fait, le coefficient de corrélation linaire de Pearson entre les variables X et Y correspond à la racine carré du coefficient de détermination, soit $r = \sqrt{R^2}$. Pour le cas des grandes surfaces, $r_{xy} = \sqrt{0,900} = 0,949$ ce que traduit une forte intensité de liaison entre l'effectif employé dans les grandes surfaces et leur chiffre d'affaires. Une corrélation significative entre le prédicteur et la variable à prédire est une condition nécessaire pour obtenir un modèle de régression linéaire simple.

4 Mesure de la signification globale du modèle linéaire

Le coefficient de détermination R^2 indique dans quelle proportion la variabilité des données est prise en compte par le modèle régressif construit ; encore faut-il que cette proportion soit significative : il faut donc vérifier la validité du modèle. Ce contrôle s'opère en deux temps : d'abord avec un test de variance pour vérifier la significativité globale du modèle linéaire (c'est l'objet de la troisième étape de la démarche) ; ensuite, en testant la significativité de chacun des coefficients de régression (objet de l'étape suivante). En effet, pour qu'un modèle de régression soit intéressant pour effectuer une prévision en toute confiance, il faut qu'il soit valide et que les estimations faites soient suffisamment précises.

La mesure de la significativité globale du modèle linéaire passe par la décomposition de la variance. Tous les logiciels statistiques calculent un tableau basé sur les sources de variabilité SCE_M et SCE_R, ainsi que les degrés de liberté (d.d.l) associés à chacune d'elles (tableau 6.4) :

– le nombre de degrés de liberté (ν) dû au modèle de régression SCE_M est égal au nombre de paramètres du modèle, sauf la constante, soit d.d.l.= 2-1=1 ;
– le nombre de degrés de liberté pour la variabilité résiduelle SCE_R est égal à n-2 et celui de la variabilité totale SCE_T vaut n-1.

Tableau 6.4 – Tableau ANOVA pour la régression

Variabilité	SCE	d.d.l.	Carré moyen	F	Probabilité critique
Due au modèle	SCE_M	1	$CM_M = SCE_M$	$F_{obs} = \frac{CM_M}{CM_R}$	*p.c.*
Résiduelle	SCE_R	n-2	$CM_R = \frac{SCE_R}{n-2}$		
Totale	SCE_T	n-1	$CM_T = \frac{SCE_T}{n-1}$		

La significativité du modèle est mesurée en comparant les carrés moyens (*CM*) dus au modèle de régression et aux résidus (et non pas leurs variabilités) en considérant le rapport des variances CM_M/CM_R. L'hypothèse nulle du test global postule que, pour la population, le prédicteur *X* n'a pas d'effet sur la variable *Y* donc $\beta_1 = 0$. La statistique utilisée pour tester de significativité globale du modèle pour un seuil d'erreur donné α est le rapport du F de Fisher-Snedecor :

$$F_{obs} = \frac{CM_M}{CM_R}$$

Sous hypothèse nulle, la statistique du test F_{obs} suit une loi de Fisher avec $v_1 = 1$ et $v_2 = n - 2$ degrés de libertés. Le test consiste alors à comparer la valeur observée F_{obs} avec la valeur critique $F_{\alpha|v_1=1|v_2=n-2}$ et de rejeter l'hypothèse nulle au profit de l'hypothèse alternative si $F_{obs} > F_{\alpha|v_1=1|v_2=n-2}$. Une autre procédure, qui rend l'interprétation du test plus directe, est de calculer la probabilité critique (*p.c.*) qui mesure l'erreur commise en rejetant l'hypothèse nulle.

Remarque

La somme des écarts des carrés due au modèle de régression SCE_M peut être faible par rapport à la somme des écarts due aux résidus SCE_R, mais pour un modèle non déterministe elle n'est jamais nulle. Le terme qui exprime les carrés moyens CM_M est toujours positif et, par conséquent $F_{obs} > 0$. Pour que le modèle ait des chances d'être significatif, il faut que la valeur observée de *F* soit nécessairement > 1 et la plus grande possible (*cf.* chapitre 2 : 7.1). La plupart de conditions d'application de la régression linéaire énumérées au début de ce chapitre sont nécessaires pour mesurer la significativité de modèle global de régression à l'aide du test paramétrique F. C'est seulement sous ces conditions que sous l'hypothèse nulle la statistique F suivra une loi de Fisher.

© Dunod – Toute reproduction non autorisée est un délit.

Application au cas BODOT. Quelle est la significativité globale du modèle de régression estimé ? Le tableau 6.5 présente les résultats numériques. La statistique F_{obs} vaut :

$$F_{obs} = \frac{CM_M}{CM_R} = \frac{16840624}{465977} = 36{,}140$$

Pour un seuil d'erreur $\alpha = 0{,}05$, avec $v_1 = 1$ et $v_2 = n - 2 = 4$, la valeur théorique obtenue à partir de la table de la loi de Fisher-Snedecor est :

$$F_{\alpha=0{,}05|v_1=1|v_2=4} = 6{,}61$$

Comme le F_{obs} est largement supérieur à la valeur théorique (45,176 > 6,61), l'hypothèse nulle peut être rejetée pour conclure que le modèle est significatif pour un seuil d'erreur inférieur à 0,05. La probabilité critique (p.c.) du test vaut 0,004, ce qui signifie que l'erreur commise en rejetant l'hypothèse nulle est seulement de quatre chances sur mille (0,4 %). Le modèle étant statistiquement significatif, la relation linéaire entre le prédicteur X et la variable prédite Y est donc validée.

L'étape suivante, qui consiste à tester la significativité des coefficients de régression β_1 et β_0, permet de vérifier si la prévision est suffisamment précise.

Tableau 6.5 – Tableau d'analyse de la variance pour la régression – cas BODOT

Variabilité	SCE	ddl	Carrés moyen	F_{obs}	p.c.
Due au modèle	16 840 624	1	16 840 624	36,140	0,004
Résiduelle	1 863 910	4	465 977		
Totale	18 704 533	5			

5 Tests de signification des coefficients de régression

Un modèle de régression n'est utile que s'il est opérationnel, c'est-à-dire que les valeurs de Y qu'il prédit sont suffisamment précises. Or, cela dépend d'abord de la significativité des coefficients de régression, significativité qu'il faut donc aussi vérifier.

Remarque

Cela se fait avec un test de Student (*cf.* § 5.1 et 5.2 ci-dessous) qui implique de respecter certaines spécifications. En effet, le test de la significativité de la pente β_1 et de la constante β_0 du modèle de régression repose sur les paramètres de la distribution d'échantillonnage des variables aléatoires b_0 et b_1. Il faut alors vérifier que b_0 et b_1 sont bien des *estimateurs sans biais* des coefficients β_0 et β_1 sur la population dont l'échantillon a été extrait. Cet aspect est développé dans la rubrique « focus » ci-après pour le lecteur intéressé.

Si la significativité des coefficients de régression est validée, cela signifie que le modèle statistique élaboré rend bien compte d'une relation linéaire entre X et Y et que les estimations faites pour les valeurs de Y sous X sont bien « expliquées » par la droite de régression.

Focus

Les caractéristiques de la distribution d'échantillonnage des coefficients de régression

Les coefficients de la droite de régression b_0 et b_1 sont des variables aléatoires dont les distributions d'échantillonnage peuvent être construites pour les échantillons de taille n. Caractériser leur distribution d'échantillonnage revient à déterminer trois paramètres : la loi de distribution, l'espérance mathématique et l'écart-standard.

Les caractéristiques de la distribution d'échantillonnage de la pente b_1

La pente de la droite de régression peut être écrite comme suit :

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

En développant le numérateur, il en résulte :

$$b_1 = \frac{\sum [(x_i - \bar{x})y_i - (x_i - \bar{x})\bar{y}]}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})y_i}{\sum (x_i - \bar{x})^2} - \frac{\bar{y}\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

La somme des écarts $\sum (x_i - \bar{x})$ étant toujours nulle, le deuxième terme ci-dessus est également nul. La pente peut alors s'écrire :

$$b_1 = \left(\frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) Y_i$$

Sachant que $\bar{x} = 1/n \sum x_i$ et que chaque x_i est constant, le terme entre parenthèses est lui aussi constant. La pente est alors une variable aléatoire car combinaison linéaire des variables aléatoires Y_i.

Loi de distribution d'échantillonnage de b_1

Puisque la pente b_1 est une combinaison linéaire des variables aléatoires Y_i, distribuées normalement par hypothèse, la distribution d'échantillonnage de b_1 est normale.

Variance de la distribution d'échantillonnage de la pente (s_{b1}^2)

$$s_{b_1}^2 = \mathrm{Var}\left(\frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} Y_i \right) = \left(\frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)^2 \mathrm{Var}(Y_i) = \frac{\mathrm{Var}(Y_i)}{\sum (x_i - \bar{x})^2}$$

La vraie variance Y_i est inconnue, mais l'estimateur sans biais de cette variance est la variance des résidus $s_R^2 = CM_R = SCE_R / (n-2)$. C'est pour cette raison que la normalité des résidus est l'une des conditions nécessaires à l'application de la régression.

En replaçant $Var(Y_i)$ par s_R^2 et en multipliant le dénominateur de l'expression ci-dessus avec le terme neutre n/n, on obtient la formule finale de la variance de la distribution d'échantillonnage de b_1:

$$s_{b_1}^2 = \frac{s_R^2}{\sum (x_i - \bar{x})^2} = \frac{s_R^2}{n s_x^2}$$

☞

© Dunod – Toute reproduction non autorisée est un délit.

L'écart-type de la distribution d'échantillonnage de b_1, appelé *écart-standard* (ou *erreur-standard*, « *standard-error* » en anglais), est donné par la formule :

$$s_{b_1} = \sqrt{\frac{s_R^2}{n s_x^2}}$$

Espérance mathématique de la distribution d'échantillonnage de b_1

$$E[b_1] = E\left(\frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} Y_i\right) = \frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} E[Y_i] = \frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} (\beta_0 + \beta_1 x_i)$$

$$E[b_1] = \frac{\beta_0 \sum(x_i - \bar{x}) + \beta_1 \sum(x_i - \bar{x}) x_i}{\sum(x_i - \bar{x})^2}$$

Sachant que $\sum(x_i - \bar{x})$ est nulle et que $\sum(x_i - \bar{x}) x_i = \sum(x_i - \bar{x})^2$, on obtient :

$$E[b_1] = \frac{\beta_1 \sum(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

Et donc

$$E[b_1] = \beta_1$$

Ce résultat est très important car il démontre que b_1 est l'estimateur sans biais de la pente inconnue β_1 et cette valeur constitue la valeur centrale de la distribution d'échantillonnage de b_1. Les trois paramètres de la distribution d'échantillonnage de b_1 étant maintenant connus, sa loi de distribution peut s'écrire :

$$b_1 \rightarrow N(\beta_1; s_{b_1}^2)$$

Les caractéristiques de la distribution d'échantillonnage de la constante b_0

Loi de distribution d'échantillonnage de b_0

La b_0 constante peut être exprimée à l'aide de l'expression suivante :

$$b_0 = \bar{Y} - b_1 \bar{x} = \frac{1}{n}\left(\sum Y_i - b_1 \sum x_i\right)$$

La variable b_0 est une variable aléatoire car fonction de la variable aléatoire b_1 distribuée normalement. De ce fait, b_0 est également distribué normalement.

Variance de distribution d'échantillonnage de b_0 ($s_{b_0}^2$)

La variance de la distribution d'échantillonnage de b_0 est :

$$s_{b_0}^2 = \text{Var}\ \bar{Y} - b_1 \bar{x}\ = \text{Var}\ \bar{Y}\ + \bar{x}^2 \text{Var}(b_1) - 2\bar{x}\text{Cov}(\bar{Y}, b_1)$$

La variance d'$\bar{Y}$ est

$$Var\left(\bar{Y}\right) = \frac{\sigma^2}{n} = \frac{\hat{s}_R}{n}$$

Il est possible de démontrer que $Cov\left(\bar{Y}, b_1\right) = 0$

☞

$$s^2_{b_0} = Var\left(\bar{Y}\right) + \bar{x}^2 Var(b_1) = \frac{s^2_R}{n} - \bar{x}^2 \frac{s^2_R}{ns^2_x}$$

On obtient ainsi la formule de la variance de la distribution d'échantillonnage de b_0

$$s^2_{b_0} = \left(\frac{1}{n} + \frac{\bar{x}^2}{ns^2_x}\right) s^2_e$$

D'où son écart-standard :

$$s_{b_0} = \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{ns^2_x}\right) s^2_e}$$

Espérance mathématique de la distribution d'échantillonnage de b_0

$$E[b_0] = E\left[\bar{Y} - b_1\bar{x}\right] = E\left[\bar{Y}\right] - \bar{x}E[b_1]$$

L'espérance mathématique de b_1 est égale à β_1 puisqu'il a été démontré que b_1 est l'estimateur sans biais de β_1. Il faut alors déterminer $E[\bar{Y}]$

$$E\left[\bar{Y}\right] = E\left[\frac{1}{n}\sum Y_i\right] = \frac{1}{n}\sum E\left[Y_i\right] = \frac{1}{n}\sum E\left[\beta_0 + \beta_1 x_i\right] = \beta_0 + \beta_1\bar{x}$$

En remplaçant les valeurs de $E[b_1]$ et de $E[\bar{Y}]$ dans la formule initiale de $E[b_0]$ on obtient l'Espérance mathématique de la distribution d'échantillonnage de b_0

$$E[b_0] = \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0$$

Ce qui signifie que b_0 est bien l'estimateur sans biais de β_0, soit la valeur centrale de la distribution d'échantillonnage de b_0 dont la loi s'écrit : $b_0 \rightarrow N(\beta_0 ; s^2_{b_0})$

Remarque

À partir de la formule de l'écart-standard du s_{b_1}, on peut observer que l'écart-standard de la distribution de la pente de la régression (s_{b_1}) diminue avec l'augmentation de la taille d'échantillon et de la variance du prédicteur.

Application au cas BODOT : calcul des deux écarts-standard :

$$s_{b_1} = \sqrt{\frac{372782}{6*81,677}} = 27,582$$

et

$$s_{b_0} = \sqrt{\left(\frac{1}{6} + \frac{18^2}{6*81,667}\right)*372782} = 555,538$$

© Dunod – Toute reproduction non autorisée est un délit.

5.1 Test et intervalle de confiance de la pente β_1

L'hypothèse nulle du test postule que le prédicteur X n'a pas d'effet statistiquement significatif sur la variable à prédire ($H_0 : \beta_1 = 0$) ; l'hypothèse alternative suppose l'inverse ($H_1 : \beta_1 \neq 0$). Si les conditions d'application sont réunies, sous H_0, donc si $\beta_1 = 0$, l'écart-réduit est :

$$t_{obs} = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1}{s_{b_1}}$$

La variance de la variable Y n'étant pas connue mais estimée et, de plus, la taille de l'échantillon étant petite, la variable aléatoire t_{obs} suit une loi de Student avec $n = n–2$ degrés de liberté. Pour un seuil d'erreur α donné, la valeur critique de la statistique $t_{\alpha|n-2}$ peut-être trouvée dans la table de Student et comparée à la valeur observée t_{obs}. La règle de décision est la suivante : si $|t_{obs}| > t_{\alpha|n-2}$, l'hypothèse nulle est rejetée. Il est toujours plus aisé de calculer la probabilité critique du test.

Dans le cas Bodot, la statistique observée vaut :

$$t_{obs} = \frac{b_1}{s_{b_1}} = \frac{185,388}{27,582} = 6,721$$

La valeur critique pour un test bilatéral au seuil d'erreur de 5 % ($\alpha = 0,05$) vaut :

$$t_{0,025|5} = 2,776$$

Puisque $|t_{obs}| > t_{\alpha|n-2}$, on rejette H_0 au bénéfice de l'hypothèse alternative pour conclure que β_1 est statistiquement significatif au seuil d'erreur de 5 %. Cela signifie que la variable Y est bien fonction linéaire de X. Si on teste β_1 au seuil d'erreur de 1 % ($\alpha = 0,01$), la valeur critique vaut $t_{0,005|4} = 4,604$, donc toujours inférieure à la valeur observée : β_1 reste statistiquement significatif à ce seuil d'erreur. La probabilité critique du test est *p.c.* = 0,004, ce qui signifie qu'il n'y a que quatre chances sur mille de se tromper en rejetant l'hypothèse nulle.

Remarque

Il est intéressant de souligner que le test de la pente de la régression est identique à celui du coefficient de corrélation linéaire de Pearson. La variable auxiliaire t_{obs} peut donc être alternativement définie comme :

$$t_{obs} = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}}$$

Illustration empirique de cette équivalence avec le **cas Bodot** où le coefficient de corrélation entre X et Y vaut :

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{15140}{9 * 1765} = 0,949$$

La valeur de t_{obs} est alors la même que celle qui vient d'être calculée ci-dessus :

$$t_{obs} = 0,949\sqrt{\frac{6-2}{1-0,949^2}} \cong 6,721$$

La décision statistique liée au test est identique à la procédure détaillée ci-dessus.

Étant donné que b_1 est l'estimateur sans biais de la vraie pente inconnue de la population, l'intervalle de confiance de la pente β_1 (IC_{β_1}) pour un seuil d'erreur α est calculé selon l'expression $b_1 - t_{\alpha/2|n-2}s_{b_1} < \beta_1 < b_1 + t_{\alpha/2|n-2}s_{b_1}$ qui peut aussi s'écrire sous la forme :

$$IC_{\beta_1} : [b_1 - t_{\alpha/2|n-2}s_{b_1}; b_1 + t_{\alpha/2|n-2}s_{b_1}]$$

Application au cas Bodot. L'intervalle de confiance de β_1 au seuil d'erreur 0,05 vaut donc :

$$IC_{\beta_1} : [185,388 - 2,571 * 27,582; 185,388 + 2,571 * 27,582]$$

$$IC_{\beta_1} : [114,475; 256,301]$$

En conséquence, pour un niveau de confiance de 95 %, la pente de régression dans la population sera comprise entre les valeurs 114,475 et 256,301. Cela signifie donc que, pour une variation de l'effectif (X) de 1 000 « personnes occupées », la valeur du CA (Y) augmente de 114 à 256 millions d'€, soit dans un rapport de 1 à 2, approximativement.

5.2 Test et intervalle de confiance de la constante β_0

Le test de significativité de la constante de la régression est très similaire à celui de la pente. L'hypothèse nulle du test postule que la constante pour la population est nulle : $H_0 : \beta_0 = 0$, versus l'hypothèse alternative $H_1 : \beta_0 \neq 0$. Si les conditions d'application sont réunies, sous H_0, donc si $\beta_0 = 0$, la statistique du test de la pente est donnée par la relation :

$$t_{obs} = \frac{b_0 - \beta_0}{s_{b_0}} = \frac{b_0}{s_{b_0}}$$

Sous H_0, la statistique du test suit la loi de Student avec $n = n–2$ degrés de liberté.

Application du test au cas BODOT

$$t_{obs} = \frac{336,354}{555,538} = 0,605$$

Dans la table de la loi de Student, pour un test bilatéral, la valeur critique de $t_{\alpha/2|n-2}$ au seuil de risque de 5 % ($\alpha = 0,05$) vaut : $t_{0,025/4} = 2,776$. Puisque t_{obs} est largement inférieur à la valeur critique $t_{\alpha/2|n-2}$, la constante n'est donc pas significative (ce résultat est logique car il n'est guère sensé de mesurer le chiffre d'affaires de magasins dont l'effectif en personnel est nul !). L'intervalle de confiance de la constate β_0 pour un niveau de confiance α et dans le cadre d'un test bilatéral est :

$$IC_{\beta_0} : [b_0 - t_{\alpha/2|n-2}s_{b_0}; b_0 + t_{\alpha/2|n-2}s_{b_0}]$$

© Dunod – Toute reproduction non autorisée est un délit.

L'intervalle de confiance de β_0 au seuil d'erreur 0,05 pour cet exemple est :

$$IC_{\beta_0} : [336,354 - 2,571 * 555,538; 336,354 + 2,571 * 555,538]$$

$$IC_{\beta_0} : [-1091,934; 1764,642]$$

En conséquence, pour un niveau de confiance de 95 %, la constante β_0 de la régression pour la population sera comprise entre les valeurs -1091,934 et 1764,642. Dans cet exemple, ces valeurs n'ont pas de sens, mais dans diverses situations, cela peut en avoir un. Par exemple, en agronomie, lors de l'étude du rendement comparatif des parcelles en fonction du dosage en engrais, la valeur de β_0 correspond au rendement (qui existe bien !) de la parcelle témoin qui n'a pas reçu d'engrais, *i.e.* dont le dosage en engrais est nul (X = 0).

5.3 Prévision pour un niveau donné de $X = x_0$

Finalement, l'objectif principal de la régression linéaire est de prévoir les valeurs de la variable à expliquer Y en fonction des valeurs de la variable explicative X. Dans cette perspective, la prévision ne sert pas à prédire une grandeur dans le futur à partir des données actuelles ou passées, mais à prédire une grandeur en fonction de ses déterminants à l'aide d'un modèle théorique linéaire. Elle repose sur une estimation des valeurs recherchées, estimation qui peut être ponctuelle ou effectuée sur la base de l'intervalle de confiance.

Estimation ponctuelle de la moyenne d'Y pour une valeur de X donnée

Les calculs jusqu'alors effectués ont permis d'estimer les coefficients de la régression et leur intervalle de confiance. Comme vu dans tout ce qui précède, le modèle linéaire simple qui vient d'être construit permet de prédire une moyenne de Y en fonction des valeurs fixes de X, ce qui correspond à une *prévision par estimation ponctuelle*. Aussi, pour prévoir la valeur moyenne de Y pour une valeur donnée de X, il suffit de remplacer une valeur du prédicteur dans le modèle linéaire $\mu_0 = \beta_0 + \beta_1 x_0$.

Application au cas Bodot. Le chiffre d'affaires pour un effectif de 26 000 personnes occupées par exemple, est de 336,354+185,388*26 = 5 156,442 millions d'euros. Comme chaque grande surface a un chiffre d'affaires plus au moins différent de celui des autres, ce chiffre d'affaires de 5 156,442 millions d'euros calculé avec le modèle de régression correspond à une estimation ponctuelle du chiffre d'affaires moyen des grandes surfaces ayant un effectif de 26 000 personnes occupées.
Compte tenu de la plus ou moins grande variabilité du réel CA des magasins ayant un tel effectif, il est alors plus pertinent de faire des prévisions de cette moyenne conditionnelle non par estimation ponctuelle, mais par intervalle de confiance. Cela repose sur la problématique de distribution d'échantillonnage déjà abordée plus haut.

Intervalle de confiance de la moyenne d'Y pour une valeur de X donnée

Pour construire l'intervalle de confiance de la moyenne, il faut d'abord déterminer son estimateur, puis son écart-standard et sa loi de distribution (*cf.* focus du § 5 plus

haut). L'estimateur de $\mu_0 = \beta_0 + \beta_1 x_0$ est donné par la droite des moindres carrés $b_0 + b_1 x_0$. Il est démontré que l'écart-standard de la distribution d'échantillonnage de cet estimateur est :

$$\sigma_{\mu_0} = \sqrt{\left(\frac{1}{n} + \frac{\left(x_k - \bar{x}\right)^2}{n s_x^2}\right) s_R^2}$$

L'intervalle de confiance de la moyenne de Y pour une valeur donnée de X est donnée par la relation :

$$IC_{\mu_0} : \left[(b_0 + b_1 x_0) - t_{\alpha/2|n-2}\sigma_{\mu_0}; (b_0 + b_1 x_0) + t_{\alpha/2|n-2}\sigma_{\mu_0}\right]$$

Calcul de cet intervalle de confiance du CA moyen appliqué au cas Bodot pour un seuil d'erreur de 5 % et pour X=26

$$\sigma_{\mu_0} = \sqrt{\left(\frac{1}{6} + \frac{(26-18)^2}{6*81,677}\right)*372782} = 332,897$$

$$IC_{\mu_0} : \left[(336,354 + 185,388*26) - 2,571*332,897; (336,354 + 185,388*26)\right.$$
$$\left.+2,571*332,897\right]$$

$$IC_{\mu_0} : [4301; 6012]$$

Ceci signifie qu'il y a 95 % des chances que la moyenne d'Y (chiffre d'affaires moyen) pour un effectif de 26 se trouve dans cet intervalle.

Intervalle de confiance des valeurs d'Y pour une valeur de X donnée

Il est également possible de construire un intervalle de confiance de façon à être sûr à $1-\alpha$ que l'ensemble des valeurs d'Y s'y trouve. Il est démontré que l'écart-standard de la distribution d'échantillonnage est :

$$\sigma_{Y_0} = \sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_k - \bar{x}\right)^2}{n s_x^2}\right) s_R^2}$$

Et l'intervalle de confiance pour une valeur unique de la variable Y en x_0 est :

$$IC_{Y_0} : \left[(b_0 + b_1 x_0) - t_{\alpha/2\,|n-2}\sigma_{Y_0}; (b_0 + b_1 x_0) + t_{\alpha/2|n-2}\sigma_{Y_0}\right]$$

Il en résulte que si la taille de l'échantillon augmente, σ_{Y_0} diminue jusqu'à donner une courbure de l'intervalle de confiance presque aplatie.

L'intervalle de confiance de la moyenne de Y est plus étroit que celui d'une valeur unique, ce qui signifie qu'il est plus difficile de prédire une valeur particulière de Y

© Dunod – Toute reproduction non autorisée est un délit.

que de prédire sa moyenne. La figure 6.4 illustre ces intervalles pour l'exemple des grandes surfaces. La droite du milieu représente la droite de régression :

$$Y = \beta_0 + \beta_1 X$$

Les deux lignes en points carrés représentent l'intervalle de confiance à 95 % de la moyenne de Y en fonction des valeurs de X, alors que les courbes en tirets délimitent l'intervalle de confiance (plus large) d'une valeur particulière de Y.

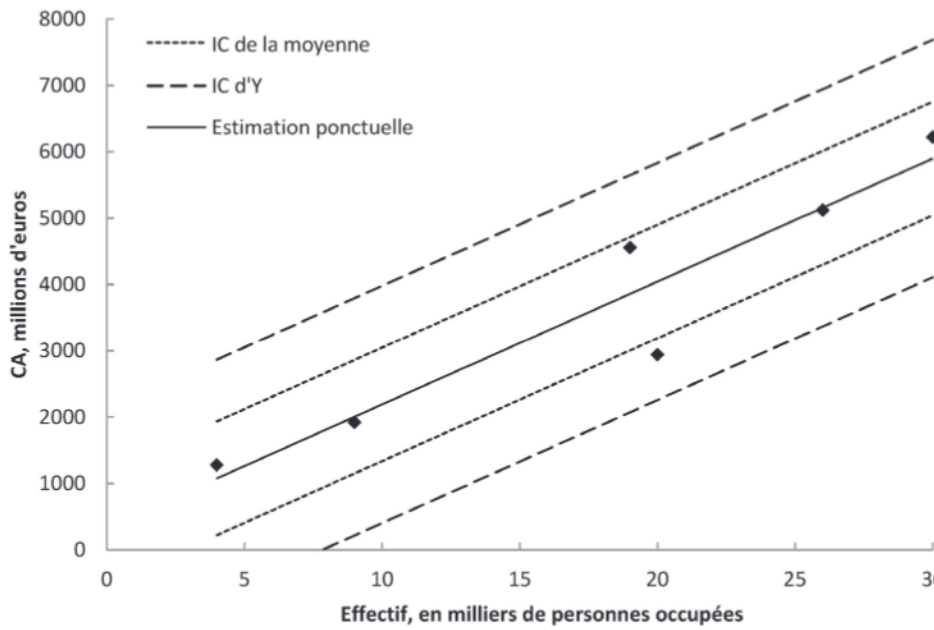


Figure 6.4 – Les intervalles de confiance à 95 % pour une valeur particulière de Y et sa moyenne

Remarque

La prévision avec un modèle linéaire composé d'un seul prédicteur donne les meilleurs résultats pour les valeurs qui sont proches de la moyenne du prédicteur. Plus qu'on s'éloigne de cette moyenne, et plus l'intervalle de confiance devient large. Notons encore que le modèle linéaire simple peut être utilisé avec une variable qualitative comprenant deux catégories (oui/non, femme/homme, etc.) comme prédicteur. Mais l'interprétation des résultats est quelque peu différente comparée à celle d'une variable quantitative.

Sur la base du modèle qu'il a construit et validé, Bodot peut donc répondre à la question qui lui a été posée. Pour le premier client (INDEP), on constate que son CA (2 800 millions €) est à l'intérieur de l'intervalle de confiance (à 95 %) de la moyenne du CA des secteurs ayant un effectif équivalent au sien [IC = 1705 ; 3417] et se trouve même légèrement supérieur au CA moyen du secteur (2561 millions €). Sa productivité semble donc satisfaisante.

Pour le second client (ENSE), on constate que son CA (3100 millions €) est *au-dessous* de la borne inférieure de l'IC à 95 % correspondant à la moyenne du CA des secteurs ayant un effectif équivalent au sien [IC = 3744 ; 5456] : en d'autres termes, le CA de ENSE se trouve nettement inférieur au CA moyen (4600) des 95 % de magasins de son secteur ; il fait donc partie des 2,5 % qui ont la moins bonne performance. Il lui faut donc s'interroger sur les raisons de cette situation (pertinence des implantations, richesse vive des zones de chalandise, motivation du personnel).

Section 2 LA RÉGRESSION LINÉAIRE MULTIPLE (RLM)

Avec la régression simple, la prévision des valeurs d'une variable dépendante (Y) repose sur une seule variable explicative (X), à l'exemple du prix d'un appartement (Y) qui dépend de sa surface (X). Mais dans la réalité, une seule variable explicative peut ne pas suffire pour effectuer une bonne prévision. C'est d'ailleurs le cas du prix d'un appartement, car il dépend aussi d'autres facteurs (quartier et emplacement, étage et ensoleillement, vétusté, etc.). Pour une bonne prévision de la variable Y, il faut s'appuyer sur plusieurs variables explicatives à la fois. C'est l'objet de la régression linéaire multiple (RLM) : elle explique les variations d'une variable dépendante (à expliquer) en fonction de plusieurs variables indépendantes (explicatives). La régression linéaire multiple a beaucoup de points communs avec la régression linéaire simple ; aussi, cette section est plutôt centrée sur ses spécificités.

Cas d'illustration

AFARC
L'attitude des Français envers le réchauffement climatique

Comprendre les attitudes du public à l'égard du changement climatique est essentiel pour la conception de politiques climatiques efficaces et pour changer le comportement individuel. Depuis plus de trente ans, de nombreux sondages d'opinion ont été réalisés sur le changement climatique global où il apparaît que le public est de plus en plus sensibilisé à la menace réelle que cela représente. De ce fait, les préoccupations liées au changement climatique ont pris une importante dimension politique.

Les nombreuses recherches – reposant sur les concepts centraux des sciences sociales et comportementales – conduites sur les attitudes du public à l'égard du ☞

© Dunod – Toute reproduction non autorisée est un délit.

changement climatique afin de le responsabiliser, mettent en évidence que l'évolution du comportement repose bien sur un certain nombre d'attitudes liées à des croyances générales sur l'environnement, à des normes, à la perception du risque environnemental dû au réchauffement climatique, et à la perception de l'action personnelle (auto-efficacité).

Dans cette perspective, une étude est conduite pour étudier l'attitude des Français adultes envers le réchauffement climatique afin d'envisager des actions concrètes en matière de communication et de formation face à cette menace. Un questionnaire de plusieurs dizaines de questions a permis de mesurer les déterminants (variables latentes) de l'attitude (Attitude) des adultes : leurs croyances sur la réalité du réchauffement climatique (Croyances), les normes personnelles ou sociales qui peuvent, à leur tour, être fonction des croyances (Normes), la conscience du risque pour les individus lié au réchauffement climatique (Risque) et la perception de leur efficacité personnelle dans la lutte contre ce phénomène (Efficacite). Sélectionnées au hasard dans la population adulte, 1 126 personnes ont répondu à ce questionnaire, complété des informations suivantes : âge (Age), revenu annuel du foyer fiscal (Revenu) et sexe (Sexe).

Les quatre premières variables et la variable Attitude sont des variables latentes : aussi, des analyses factorielles en composantes principales ont été réalisées pour construire les variables finales (facteurs). Le tableau 6.6 fournit les variables finales impliquées dans les calculs de la régression multiple. Toutes ces variables, excepté la variable Sexe, sont de variables métriques.

Tableau 6.6 – Les variables impliquées dans la régression multiple

Observations	Attitudes	Normes	Risque	Croyances	Efficacité	Age	Sexe	Revenu
1	1,39	2,18	4,59	1,51	3,83	22	Homme	5 000
2	2,87	1,65	4,59	1,94	3,56	23	Homme	15 000
3	1,26	1,66	4,59	2,37	3,51	37	Homme	23 000
4	0,78	1,29	4,59	0,91	3,47	24	Homme	39 000
5	3,09	3,63	4,59	3,71	3,65	21	Femme	61 000
…	…	…	…	…	…	…	…	…
1 124	5,13	5,01	8,6	5,04	3,94	22	Femme	118 000
1 125	4,40	3,76	8,6	5,04	3,78	58	Homme	140 000
1 126	4,92	4,67	8,6	4,67	3,86	37	Femme	143 000
Moyenne	*4,043*	*3,970*	*7,341*	*4,111*	3,762	36,513		*56 283,304*
Écart-type	*0,795*	*0,810*	*0,844*	*0,790*	*0,086*	*15,482*		*28 634,472*

L'un des objectifs de cette étude – celui qui est spécifiquement traité dans cet exemple – est de vérifier si l'attitude envers le réchauffement climatique peut être expliquée, comme suggéré dans certaines recherches, par les croyances personnelles, les normes partagées, la perception du risque et par sa propre perception de l'efficacité individuelle. Si cette relation est au moins partiellement sinon totalement avérée, cela justifie d'entreprendre des actions de formation et de communication ciblées dont on pourra vérifier ensuite l'efficience auprès du public.

Pour vérifier statistiquement cette relation, il faut alors examiner si l'« attitude envers le réchauffement climatique » des Français adultes peut être « expliquée » par une combinaison linéaire des quatre autres variables.

Focus

Conditions d'application de la RLM

À l'instar de la régression linéaire simple, la régression linéaire multiple s'applique sous certaines conditions résumées ci-après :

- Toutes les variables doivent être quantitatives. Des variables nominales peuvent être utilisées, mais elles doivent d'abord être converties en variables muettes ;
- Les résidus ε_i sont distribués normalement, d'espérance nulle et de variance σ^2 (homoscédasticité). Cette variance est souvent inconnue et doit être remplacée par son estimateur sans biais ;
- Les résidus ε_i doivent être indépendants : $Cov(\varepsilon_i, \varepsilon_j) = 0$ $(\forall\ i \neq j)$;
- Les relations entre la variable à expliquer Y et les variables explicatives prises une à une doivent être linéaires. Si cette relation n'est pas linéaire, il est possible de transformer les variables explicatives mais l'interprétation des résultats et des tests devient alors difficile ;
- Les variables explicatives ne doivent pas être corrélées fortement entre elles (colinéarité aussi réduite que possible) ;
- Les erreurs doivent être indépendantes des variables explicatives.

1 Principe du modèle RLM et structure des données

L'objectif de la RLM est d'établir une relation linéaire entre la variable Y et les p variables explicatives. Pour une population entière, la dépendance linéaire prend la forme suivante :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Mais comme dans la plupart des cas, les caractéristiques de la population entière ne peuvent pas être mesurées, on utilise un échantillon de taille n pour estimer les $p + 1$ coefficients de régression β. La structure des données de la RLM se présente

© Dunod – Toute reproduction non autorisée est un délit.

de manière comparable à celle de la RLS, sous la forme d'un tableau composé de n lignes et de $p + 1$ colonnes (Tableau 6.7). Chaque ligne représente une observation alors que les colonnes représentent les variables explicatives $(x_1, x_2,\ldots x_p)$ de sorte que $p \leq n$, plus la variable à expliquer.

Tableau 6.7 – La structure des données soumises à la régression linéaire multiple.

Observations	X_1	X_2	…	X_j	…	X_p	Y
1	x_{11}	x_{12}	…	x_{1j}	…	x_{1p}	y_1
2	x_{11}	x_{22}	…	x_{2j}	…	x_{2p}	y_2
…	…	…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ip}	y_i
…	…	…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{np}	y_n
Moyennes	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_j$	…	$\bar{x}_p$	$\bar{y}$
Écarts-types	s_1	s_2	…	s_j	…	s_p	y_n

La démarche pour résoudre une problématique statistique avec la régression linéaire multiple est très semblable à celle de la régression linéaire simple :

– estimer les $p + 1$ coefficients de régression à partir des données observées ;
– mesurer le pouvoir explicatif du modèle ;
– mesurer le niveau global de signification du modèle linéaire ;
– évaluer globalement ou individuellement le rôle des variables prédictives dans le modèle (significativité des coefficients de régression) ;
– évaluer la qualité du modèle lors de la prévision (intervalle de confiance) ;
– détecter les observations qui peuvent fortement influencer le résultat (points atypiques) ;
– analyser la colinéarité des variables explicatives ;
– prendre la décision.

Le « Focus » ci-dessous développe le principe du modèle de la RLM.

Focus

Principe du modèle de la RLM

À l'instar de la régression linéaire simple, la régression linéaire multiple s'applique sous certaines conditions résumées ci-après :

Pour deux variables explicatives, la distribution des observations est graphiquement représentée sur un plan de régression comme le montre la figure 6.5. Soit un échantillon de taille n qui contient une variable à expliquer Y et deux variables explicatives X_1 et X_2 (p = 2) : les données se présentent sous la forme de n triplés de valeurs :

$$(x_{11}, x_{21}, y_1),(x_{12}, x_{22}, y_2),\ldots,(x_{1i}, x_{2i}, y_i),\ldots,(x_{1n}, x_{2n}, y_n)$$

☞

La valeur observée y_i s'exprime alors comme suit :

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$$

Comme la variable Y est une variable aléatoire, il convient alors de considérer que la valeur Y_i n'est pas une valeur fixe mais plutôt la moyenne de la distribution de la variable *Y*. Il s'agit donc de $E(y_i)$ qui se trouve dans le plan d'équation correspondant au modèle linéaire à deux variables explicatives dont l'équation est :

$$E(y_i) = \hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

La valeur observée y_i peut être exprimée comme la somme de la valeur prédite par le modèle de régression plus l'erreur aléatoire : $y_i = \hat{y}_i + e_1$

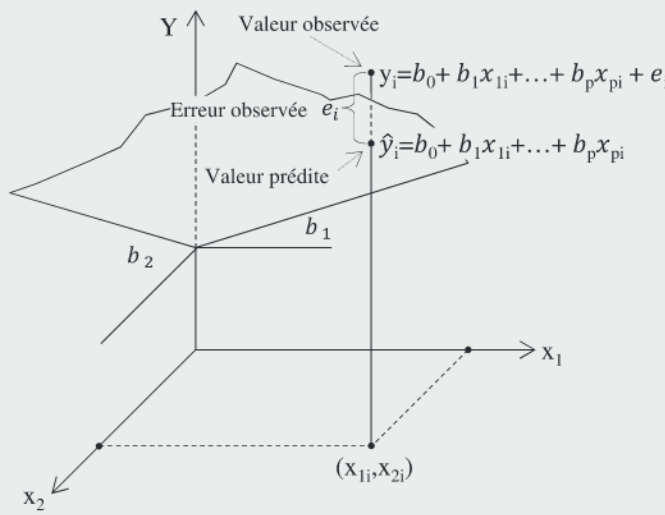


Figure 6.5 – Dispersion des observations : cas de deux variables explicatives

Le plan est ajusté selon le principe des moindres carrés où les sommes des carrés des erreurs d'estimation de la variable dépendante sont minimisées. Il est possible d'étendre ce raisonnement avec plus de deux variables explicatives.

Estimation des coefficients de régression à partir des données observées

Les *n* valeurs prises par la variable dépendante Y $(y_1, y_2, \ldots, y_n)$ peuvent être exprimées avec les équations suivantes :

$$y_1 = b_0 + b_1 x_{11} + b_2 x_{21} + \ldots + b_p x_{p1} + e_1$$

$$y_2 = b_0 + b_1 x_{12} + b_2 x_{22} + \ldots + b_p x_{p2} + e_2$$

$$\ldots$$

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi} + e_i$$

$$\ldots$$

$$y_n = b_0 + b_1 x_{1n} + b_2 x_{2n} + \ldots + b_p x_{pn} + e_n$$

© Dunod – Toute reproduction non autorisée est un délit.

☞

Il s'agit alors de calculer les coefficients $b_0, b_1, \ldots, b_p$ tels que la somme des carrés des résidus soit la plus faible possible. La solution repose sur les calculs matriciels nécessaires pour résoudre un tel problème, fondé sur un système de équations linéaires de $p + 1$ inconnues. Il y a trois cas de figure :

– il y a *autant de données que de paramètres* ($n = p + 1$) : il est possible de trouver un modèle qui s'ajuste parfaitement (la somme des carrés des résidus est nulle). C'est le cas le plus fréquent, surtout quand on cherche à minimiser le nombre de mesures pour des raisons de coût liés à l'étude.
– il y a *plus de données que de paramètres* ($n > p + 1$) : il est possible d'obtenir un modèle tel que la somme des carrés des résidus soit la plus faible possible. Pour ces deux premiers cas, une solution est possible si les variables vecteurs « colonne » ne sont pas linéairement combinées entre elles.
– il y a *moins de données que de paramètres* ($n < p + 1$) : dans ce cas, il y a une infinité de solutions possibles qui s'ajustent parfaitement aux données. Mais, dans la pratique, cette configuration est plutôt rare.

Il est compliqué d'effectuer manuellement le calcul des coefficients de régression multiple, surtout quand le nombre de variables explicatives est élevé, mais il n'est pas indispensable de connaître les procédures de calculs matriciels, les logiciels de traitement statistique s'en chargeant très bien !

Les pentes de régression, calculées à partir des données observées, sont des coefficients de régression non standardisés. Ils expriment la contribution unique d'une variable explicative X_i pour rendre compte des valeurs de la variable à expliquer Y et dépendent des unités de mesure des variables explicatives qui peuvent être très différentes. Le recours à des coefficients standardisés permet d'éliminer l'effet des unités de mesure. Pour calculer les coefficients de régression multiple standardisés, toutes les variables (même la variable à expliquer) sont transformées en variables centrées-réduites. Le calcul de ces coefficients utilise l'unité d'écart-type et non l'unité de mesure des variables. À titre d'exemple, un coefficient standardisé de 0,456 signifie qu'un accroissement de la valeur de la variable explicative d'un écart-type se traduit par une augmentation de celle de la variable à expliquer de 0,456 écart-type. Les coefficients standardisés varient entre -1 et 1 et contrairement aux coefficients non standardisés b_i, ils permettent de comparer l'importance de l'impact des variables explicatives sur la variable à expliquer.

Illustration avec le cas AFARC

Calculés à partir des données recueillies dans l'enquête réalisée auprès des 1 126 adultes, le tableau 6.8 fournit les coefficients non-standardisés avec leur écart-standard, et les coefficients standardisés. Dans ce modèle RLM, la constante b_0 indique quelle serait la valeur prédite pour l'ATTITUDE si toutes les variables explicatives étaient nulles, soit -5,837 (ce qui, dans cette hypothèse, n'a guère de signification).

Pour analyser l'effet de l'une des six variables prédictives dans la prévision des valeurs de l'ATTITUDE, on dispose de trois valeurs, par exemple pour la variable RISQUE : le coefficient non standardisé (0,203), le coefficient standardisé (0,025) et l'écart-standard (0,025). L'examen des six coefficients standardisés laisse à penser que les NORMES per-

sonnelles et collectives, les CROYANCES et la perception de l'EFFICACITÉ PERSONNELLE sur l'environnement sont les meilleurs prédicteurs de l'attitude des Français envers le réchauffement climatique, car ils ont les coefficients standardisés (Bêta) les plus élevés. À considérer la faible valeur du bêta (proche de 0) des trois autres facteurs en comparaison des trois précédents, il semble bien qu'ils aient une influence moindre sur la prédiction des valeurs de cette attitude globale.

Tableau 6.8 – Résultats de calculs matriciels rendus par les logiciels statistiques

	Coefficients non standardisés		Coefficients standardisés
	b_i	Ecart-standard (S_{b_i})	Bêta
Constante	-5,837	0,757	
NORMES	0,268	0,026	**0,273**
RISQUE	0,203	0,025	0,025
CROYANCES	0,260	0,028	**0,259**
EFFICACITÉ	1,665	0,216	**0,179**
AGE	-0,001	0,001	-0,023
REVENU (Rev)	5,856E-7	0,000	0,021

Le modèle formel initial $Y=b_0 + b_1N + b_2R + b_3C + b_4EFF + b_5A + b_6Rev$ prend donc les valeurs estimées suivantes : $Y = -5,837 + 0,268*N + 0,203*R + 0,260*C + 1,665*EFF - 0,001*A + 5,856^{e-7}*Rev$

Mais pour l'utiliser en toute confiance et pouvoir décider des thèmes prioritaires sur lesquels faire porter les actions de communication et de formation, encore faut-il être assuré de sa significativité et de sa validité.

2 Mesure du pouvoir explicatif du modèle

Le calcul des $p+1$ coefficients b_0 et b_1 a permis d'ajuster du mieux possible les valeurs initiales sur la droite de régression pour faire en sorte que la somme des carrés des erreurs soit minimale. Mais quelle est la qualité de cet ajustement, c'est-à-dire : les données multidimensionnelles observées (concernant six variables) sont-elles *fidèlement* et bien représentées sur la droite de régression ? Pour le vérifier, le raisonnement basé sur la décomposition de la variance (voir « focus » ci-dessous) est identique à celui rencontré pour la régression linéaire simple.

© Dunod – Toute reproduction non autorisée est un délit.

Focus

Décomposition de la variance

L'écart total de chaque valeur observée y_i à la moyenne d'Y $(y_i - \bar{y})$ pour chacune des n observations, peut être exprimé comme étant la somme d'une part de l'écart de la valeur prédite par le modèle $\hat{y}_i$ à la moyenne d'Y $(\hat{y} - \bar{y})$ et, d'autre part de l'écart de la valeur observée y_i à la valeur prédite par le modèle $\hat{y}_i$. Cette dernière est calculée à partir de l'équation : $\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i} + \ldots + b_px_{pi}$.

Nous obtenons alors la relation qui exprime la variabilité totale de la variable Y autour de sa moyenne : $SCE_T = SCE_M + SCE_R$. Le tableau 6.9 illustre ces calculs pour le cas AFARC.

La composante SCE_M exprime la part de la variabilité expliquée par le modèle de régression. Plus le terme SCE_M est grand et le terme SCE_R est petit, plus la droite de régression représentera fidèlement le « nuage de points ». Le rapport SCE_M/SCE_T donne alors une bonne mesure de la proportion de la variation totale expliquée par le modèle régression. Comme pour la RLS, ce rapport est appelé *coefficient de détermination multiple* et noté R^2 :

$$R^2 = \frac{SCE_M}{SCE_T} = \frac{SCE_T - SCE_R}{SCE_T} = 1 - \frac{SCE_R}{SCE_T}$$

Comme déjà vu pour la RLS (*cf.* fin du §3, section 1), le coefficient de détermination multiple R^2 est l'indicateur qui permet de juger la qualité de la régression linéaire multiple : il mesure l'adéquation entre le modèle de régression et les données observées.

On peut en déduire la valeur du coefficient de corrélation multiple, lequel exprime la corrélation entre les valeurs observées y_i et celles prédites par le modèle $\hat{y}_i$. Le coefficient de corrélation multiple correspond à la racine carrée du coefficient de détermination multiple. Pour faire la distinction avec le coefficient de corrélation bivarié, il est noté avec la lettre majuscule $R = \sqrt{R^2}$. Étant donné que $0 \leq R^2 \leq 1$, le coefficient de corrélation multiple ne peut pas avoir de valeurs négatives et varie donc entre 0 et 1 $(0 \leq R \leq 1)$.

Tableau 6.9 – Bilan des variations des données observées

Observations	*ATTITUDE*	$\hat{y}$	$(y_i - \bar{y})^2$	$(y_i - \hat{y})^2$	$(\hat{y} - \bar{y})^2$
1	1,39	2,429	7,037	1,081	2,602
2	2,87	1,955	1,375	0,838	4,360
3	1,26	1,976	7,743	0,513	4,269
4	0,78	1,453	10,645	0,454	6,704
5	3,09	3,124	0,908	0,001	0,844
…	…	…	…	…	…
1 124	5,13	5,169	1,182	0,002	1,269
1 125	4,40	4,545	0,128	0,021	0,252
1 126	4,92	4,848	0,770	0,005	0,649
Somme	***4 552,06***		***711,028***	***320,114***	***390,914***
Moyenne	***4,043***		SCE_T	SCE_R	SCE_M
Écart-type	***0,795***				

Application au cas AFARC

Le pouvoir explicatif du modèle de régression est calculé avec le coefficient de détermination qui vaut :

$$R^2 = 1 - \frac{320,114}{711,028} = 0,550$$

et

$$R^2_{ajusté} = 1 - \frac{(1-0,550)(1126-1)}{1126-6-1} = 0,547$$

Ce qui signifie que le modèle calculé « explique » 54,7 % de la variabilité des valeurs observées sur l'échantillon étudié. Le coefficient de corrélation multiple vaut $R = \sqrt{0,550} = 0,741$.

3 Mesure de la signification globale du modèle de régression multiple

Le principe a déjà été exposé dans la présentation de la RLS (§ 4) : c'est le test de Fisher-Snedecor qui considère le rapport entre la variance des écarts qui est expliquée par le modèle à la variance des résidus. Étant inconnue dans la plupart des cas, la variance de la population σ^2 doit donc être estimée à partir des écarts résiduels entre Y et le modèle. La variance s_R^2 des écarts résiduels s'écrit :

$$s_R^2 = CM_R = \frac{SCE_R}{n-p-1} = \frac{\sum e_i^2}{n-p-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}$$

© Dunod – Toute reproduction non autorisée est un délit.

Avec : p - nombre des variables explicatives, n - taille de l'échantillon et $n - p - 1$, le nombre de degré de liberté. La variance des résidus ainsi définie est un estimateur non-biaisé de la variance de la population. L'écart-type des résidus est alors donnée par la formule :

$$s_R = \sqrt{\frac{SCE_R}{n-p-1}}$$

La variance des écarts qui est expliquée par le modèle de régression (carré moyen) est égale à la somme des écarts au carré divisé par le nombre de degré de liberté qui est toujours égal au nombre de variables explicatives :

$$s_M^2 = CM_M = \frac{SCE_M}{p}$$

L'écart-type est alors :

$$s_M = \sqrt{\frac{SCE_M}{p}}$$

Le tableau 6.10 résume les calculs des sommes des écarts au carré : SCE_T, SCE_M et SCE_R et les variances associées (carrés moyens).

Tableau 6.10 – Calcul des écarts et des variances associées

Variabilité	Somme des carrés	d.d.l.	Carrés moyens
Expliquée par la régression	$SCE_M = \sum(\hat{y}_i - \bar{y})^2$	p	$S_M^2 = \frac{SCE_M}{p}$
Non expliquée (résiduelle)	$SCE_R = \sum(y_i - \hat{y}_i)^2$	n – p – 1	$s_e^2 = \frac{SCE_R}{n-p-1}$
Variation totale	$SCE_T = \sum(y_i - \bar{y})^2$	n – 1	

Dans le test du F de Fisher-Snedecor, l'hypothèse nulle postule que tous les coefficients de régression (pentes) sont nuls, donc :

$$H_O : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

Si l'hypothèse nulle est vraie, aucune des variables explicatives n'est significative dans le modèle. L'hypothèse alternative H_1 postule qu'au moins un des coefficients de régression n'est pas nul. La statistique F du test peut être exprimée par l'une ou l'autre de ces deux expressions qui sont équivalentes :

$$F_{obs} = \frac{CM_M}{CM_R} = \frac{n-p-1}{p} x \frac{SCE_M}{SCE_R}$$

ou bien

$$F_{obs} = \frac{n-p-1}{p} x \frac{R^2}{1-R^2}$$

Sous hypothèse nulle, la variable aléatoire F suit la loi de Fisher-Scedecor à $v_1 = p$, $v_2 = n - p - 1$ degrés de liberté. La règle de décision reste toujours la même : si la valeur de la statistique F observée est supérieure à la valeur critique $F_{\alpha|p|n-p-1}$, le modèle de régression est significatif au seuil d'erreur α.

Illustration avec le cas AFARC :

Pour obtenir la valeur de la variable F_{obs}, il faut commencer par calculer la variance des résidus s_R^2 et la variance du modèle s_M^2 :

$$s_R^2 = \frac{320,114}{1126-6-1} = 0,286 \; ; \; s_M^2 = \frac{390,914}{6} = 65,152$$

La statistique observée du test vaut alors : $F_{obs} = \frac{65,152}{0,286} = 227,748$

Le calcul avec l'expression alternative du donne le même résultat :

$$F_{obs} = \frac{1226-6-1}{6} * \frac{390,914}{320,114} = 227,748$$

Au risque α de 5 %, le seuil théorique vaut : $F_{0,05|4|15} = 2,106$

La valeur du F observée étant largement supérieure à la valeur critique (227,48 > 2,106), cela permet de rejeter l'hypothèse nulle et de conclure qu'il y a *au moins un des coefficients* de régression qui, de manière significative, n'est pas nul.

Les logiciels de traitement statistique résument ces résultats sous la forme d'un tableau d'analyse de variance (Anova) tel que présenté dans le tableau 6.11.

Tableau 6.11 – Tableau des résultats de l'Anova

Source des écarts	SCE	ddl	Variance	F	Sig.
Régression	390,914	6	65,152	227,749	0,000
Résidus	320,114	1 119	0,286		
Total	711,028	1 125			

En dépit d'une qualité globale, parfois très élevée, du modèle de régression, il est fréquent que certaines variables explicatives n'aient pas d'impact significatif sur la variable à expliquer. Comme l'un des objectifs de la régression linéaire multiple est aussi de construire un modèle qui soit parcimonieux, de ce fait il est possible de ne garder dans le modèle final que les seules variables explicatives qui ont un impact réellement significatif sur la variable à expliquer. Mais pour ce faire, il faut tester la significativité statistique de chacun des $p + 1$ coefficients de régression.

© Dunod – Toute reproduction non autorisée est un délit.

4 Test de la significativité des coefficients de régression

La démarche pour tester la significativité des $p + 1$ coefficients de régression multiple est identique à celle exposée pour la RLS. Les paramètres nécessaires pour ce test ont déjà été calculés avec la droite des moindres carrés, à savoir : les estimateurs des vrais coefficients pour la population (b_0, b_1, b_2,…, b_p), et les écarts-standards (s_{b_0}, s_{b_1}, s_{b_2},…, s_{b_p}).

Pour chaque coefficient à tester, l'hypothèse nulle postule que $\beta_j = 0$, alors que l'hypothèse alternative postule que $\beta_j \neq 0$. Si les conditions d'application de la régression linéaire multiple sont réunies la statistique du test est :

$$t_{b_j} = \frac{b_j}{s_{b_j}} \text{ où } (j = 0,1,2,\ldots,p)$$

qui suit la loi de Student avec $v = n - p - 1$ degrés de libertés.

Pour tester les coefficients de régression un à un, pour un seuil d'erreur α donné, il faut d'abord calculer la valeur observée du t_{b_j} et la comparer ensuite à la valeur critique $t_{\alpha/2|n-p-1}$ de la table de Student, laquelle sera d'ailleurs identique pour tous les tests individuels. La règle de décision est toujours la même : si pour un niveau de confiance donné, la valeur absolue de la statistique observée est supérieure à la valeur critique $|t_{b_j}| > t_{\alpha/2|n-p-1}$, on rejette l'hypothèse nulle au bénéfice de l'hypothèse alternative.

Application au cas AFARC. À titre d'exemple, calculons la statistique t pour la variable NORMES :

$$t_{Normes} = \frac{b_{NOR}}{s_{b_{NOR}}} = \frac{0,268}{0,026} = 10,308$$

Cette statistique suit une loi de Student à $n - p - 1 = 1126 - 6 - 1 = 1119$ degrés de liberté. Au risque α de 5 %, le seuil critique $t_{0,05|1119}$ vaut 2,244[1], et la probabilité critique pour la valeur calculée de t_{Normes} est très faible (p.c < 0,000). La valeur observée étant supérieure à la valeur théorique, l'hypothèse nulle est rejetée pour conclure qu'au seuil d'erreur de 5 %, la variable explicative NORMES a bien un effet significatif sur l'ATTITUDE. Elle sera donc incluse dans le modèle final.

À l'instar de la RLS, il est aussi possible de déterminer l'intervalle de confiance des coefficients de régression pour un seuil d'erreur donné. L'intervalle de confiance de la pente β pour un niveau de confiance α est :

$$IC_{\beta_j} : [b_j - t_{\alpha/2|n-p-1} s_{b_j} ; b_j + t_{\alpha/2|n-p-1} s_{b_j}]$$

1. Dans le cas d'un échantillon très grand, comme dans le cas AFARC, la distribution de la Loi de Student est très proche de la loi normale centrée réduite : c'est pourquoi certains logiciels utilisent 1,96 (au lieu du 2,244 calculé).

L'intervalle de confiance à 95 % de la pente relative à la variable NORMES vaut :

$$IC_{\beta_{Normes}} : [0,268 - 1,96 * 0,026; 0,268 - 1,96 * 0,026]$$

$$IC_{\beta_{Normes}} : [0,217; 0,320]$$

Ce qui veut dire qu'au niveau de confiance de 95 %, le coefficient régression β_{Normes} pour la population sera compris entre les valeurs 0,217 et 0,320.

La procédure de test qui vient d'être décrite peut être utilisée pour estimer la significativité de chacun de coefficients de régression et leur intervalle de confiance, pour un seuil d'erreur. La procédure de calcul étant identique pour tous les coefficients, le tableau 6.12 affiche les résultats pour tous ceux du cas AFARC.

Tableau 6.12 – Significativité des coefficients de régression et leurs intervalles de confiance : modèle avec six variables explicatives

Variables	Coefficients non standardisés		Coefficients standardisés	t_i	p.c..	Intervalle de confiance à 95 %	
	b_i	s_{b_i}				Borne inf.	Borne sup.
(Constante)	-29,381	3,793		-7,746	0,000	-36,823	-21,939
Normes	0,268	0,026	0,273	10,226	0,000	0,217	0,320
Risque	0,203	0,025	0,215	8,169	0,000	0,154	0,252
Croyances	0,260	0,028	0,259	9,196	0,000	0,205	0,316
Âge	-0,001	0,001	-0,023	-1,130	**0,259**	-0,003	0,001
Revenu	0,000	0,000	0,021	1,050	**0,294**	0,000	0,000
Efficacité	8,099	1,049	0,179	7,724	0,000	6,042	10,157

Le tableau 6.12 permet d'identifier facilement les variables qui sont significatives en examinant la significativité (colonne *p.c.*) du *t*-test : ainsi les coefficients relatifs aux variables *Âge* et *Revenu* ne sont pas significatifs puisque leur seuil d'erreur est beaucoup trop élevé (respectivement 25,9 et 29,4 %) par rapport au seuil d'acceptabilité de 5 %. Bien que la qualité globale du modèle de régression calculé avec les 6 variables soit bonne, deux variables explicatives peuvent être ignorées sans trop détériorer la qualité du modèle.

Le tableau 6.13 produit les coefficients du modèle composé des quatre seules variables explicatives, avec leurs caractéristiques statistiques. La qualité de ce modèle ($R^2 = 0,549$) est équivalente à celle du modèle initial comprenant 6 variables ($R^2 = 0,550$) mais avec seulement 4 variables, il est plus économique. Ainsi, les variables AGE et REVENU ne sont responsables que de 55,0 % - 54,9 % = 0,1 % de la variabilité totale de la variable ATTITUDE à expliquer !

© Dunod – Toute reproduction non autorisée est un délit.

Tableau 6.13 – Significativité des coefficients de régression et leurs intervalles de confiance : modèle avec quatre variables explicatives

Variables	Coefficients non standardisés		Coefficients standardisés	t_i	p.c..	Intervalle de confiance à 95 %	
	b_i	s_{b_i}				Borne inf.	Borne sup.
(Constante)	-29,256	3,789		-7,721	0,000	-36,691	-21,821
Normes	0,267	0,026	0,272	10,202	0,000	0,216	0,319
Risque	0,206	0,025	0,218	8,308	0,000	0,157	0,254
Croyances	0,259	0,028	0,257	9,166	0,000	0,204	0,314
Efficacité	8,059	1,048	0,178	7,690	0,000	6,003	10,116

Ce tableau de résultats montre aussi l'importance qu'il y a de travailler sur les coefficients standardisés. La comparaison entre les deux est révélatrice des erreurs de jugement qui pourraient survenir : par exemple, le coefficient non standardisé $b_{Efficacité}$ (8,059) a la valeur absolue la plus importante, alors même que sa valeur standardisée (0,178) indique que c'est cette variable EFFICACITÉ qui a le moins d'effet sur la variable à expliquer (bien que son effet soit réel et significatif). Il existe des procédures de régression pas à pas qui permettent de sélectionner les variables influentes et de les hiérarchiser selon l'importance de leur effet sur les variations de la variable à expliquer.

Application managériale au cas AFARC. Compte tenu du plan de formation et d'information programmé pour influer sur l'attitude générale du public envers le réchauffement climatique, cette étude montre que les quatre déterminants identifiés sont pertinents et bien liés à l'ATTITUDE GÉNÉRALE. Mais deux d'entre eux semblent avoir une incidence particulière : les CROYANCES et les NORMES. C'est donc sur ces deux aspects que les actions de sensibilisation (information et formation) doivent se focaliser en premier lieu, car la contribution des deux autres déterminants à l'évolution de l'attitude est beaucoup moins forte, ce que confirme la régression pas à pas réalisée.

5 Identifier le modèle le plus parcimonieux

Après avoir extrait la ou les variables « explicatives » qui ne sont pas significatives, la significativité des autres variables peut changer, à tel point que l'une d'elles, alors non significative, peut le devenir. Deux questions se posent alors : pourquoi le retrait ou l'ajout d'une variable peut modifier la significativité des autres ? Et comment construire le modèle le plus parcimonieux qui soit aussi le plus « efficace » ?

Si la solution est simple à expliquer, elle est beaucoup plus difficile à mettre en œuvre, surtout quand le nombre des variables explicatives est élevé, car il faut tester toutes les combinaisons possibles entre les p variables explicatives avec la variable

à expliquer et recalculer à chaque fois la qualité du modèle. Devant la quasi-impossibilité pratique de réaliser tous les calculs nécessaires, il existe plusieurs procédures heuristiques, proposées par la plupart des logiciels, qui ont été développées pour choisir les meilleurs modèles régression. Par exemple, le Logiciel SPSS propose cinq méthodes de sélection des variables :

– *Introduire (*Méthode *Enter).* La procédure consiste à introduire toutes les variables d'un bloc en une seule étape ; cette méthode est efficace s'il n'y a pas de multicolinéarité (*cf.* § 7) entre les variables explicatives.
– Sélection ascendante ou *Introduction progressive des variables (*Méthode *Forward).* Cette procédure itérative consiste à introduire les variables une par une pour le calcul du modèle. Après avoir fixé un critère d'introduction (généralement une valeur du *F* ou du *t*), la variable explicative qui contribue le plus et de manière la plus significative à une augmentation de la somme des carrés du modèle (augmentation du R^2) est introduite en premier, mais seulement si elle satisfait le critère d'introduction fixé ; il est procédé de même avec les suivantes. La procédure prend fin quand aucune variable restante ne satisfait le critère d'introduction.
– É*limination descendante des variables (*Méthode *Backward* ou *Précédent).* Cette procédure consiste à intégrer l'ensemble des variables explicatives dans le calcul du modèle, pour ensuite enlever celles qui ne sont pas significatives. Est éliminée en premier la variable ayant la plus petite corrélation partielle avec la variable à expliquer (ou le plus petit *t* de Student en valeur absolue). La procédure se termine quand plus aucune variable explicative ne satisfait aux critères d'élimination. Le nombre de modèles à calculer ne dépasse pas le nombre des variables explicatives.
– *Régression pas-à-pas (*Méthode *Stepwise* ou *Étape par étape).* Il s'agit d'une procédure qui combine les deux précédentes. La procédure consiste à introduire une variable explicative ayant la plus petite probabilité de F. De l'autre côté, les variables déjà incluses dans le modèle sont exclues si leur probabilité de F devient trop grande. Les seuils de probabilité du F pour l'introduction ou le retrait d'une variable est de 0,15. Cette méthode est utile en cas de multicolinéarité entre les va*riables explicatives.*
– *Élimination en bloc (*Méthode *Éliminer bloc).* Toutes les variables non significatives d'un bloc sont supprimées en une seule étape. Cette méthode est utilisée quand il est nécessaire de déterminer l'ordre d'entrée des variables dans le modèle à l'aide de blocs de variables qui seront introduits de manière hiérarchisée dans le modèle de régression (régression hiérarchique, *hierarchical regression*)

La plupart des logiciels de traitement statistique d'aujourd'hui offre une procédure de manière automatique. À titre d'exemple, pour le cas AFARC, le tableau 6.14 montre la qualité des quatre modèles régressifs calculés avec le logiciel SPSS avec la procédure *stepwise* (*Pas à pas*). Le meilleur modèle intègre quatre variables explicatives (54,7 % de la variance totale expliquée) lesquelles correspondent bien à celles identifiées comme significatives dans la partie précédente.

© Dunod – Toute reproduction non autorisée est un délit.

Table 6.14 – Récapitulatif des modèles : SPSS, procédure étape par étape

Modèle	R	R^2	R^2 ajusté	Erreur standard de l'estimation	Modifier les statistiques				
					Variation de R-deux	Variation de F	ddl1	ddl2	Sig. Variation de F
1	,631[a]	,399	,398	,61679	,399	745,020	1	1124	,000
2	,696[b]	,484	,484	,57132	,086	187,050	1	1123	,000
3	,725[c]	,525	,524	,54861	,041	95,856	1	1122	,000
4	,741[d]	,549	,547	,53493	,024	59,143	1	1121	,000

a. Prédicteurs : (Constante), Croyances

b. Prédicteurs : (Constante), Croyances, Normes

c. Prédicteurs : (Constante), Croyances, Normes, Risque

d. Prédicteurs : (Constante), Croyances, Normes, Risque, Efficacité

e. Variable dépendante : Attitude

6 Mesure de la colinéarité entre les variables explicatives

La colinéarité correspond au fait que des variables sont fortement liées entre elles. Or, comme déjà vu, l'une des conditions d'application de la régression linéaire multiple est l'indépendance des variables explicatives. Si elles sont fortement liées, voire en relation parfaite, cela pose deux problèmes :

– Le premier est d'ordre statistique. La matrice des coefficients aura des colonnes résultant d'une combinaison linéaire entre variables, et dans ce cas, la matrice inverse n'est pas calculable : il n'y a donc pas de solution.
– Le second problème est d'ordre logique. Une variable qui est linéairement combinée avec une autre déjà présente dans le calcul n'apporte aucune information supplémentaire aux données : elle est donc inutile.

Si dans le champ de la gestion, il est plutôt rare de trouver des variables très fortement liées (colinéaires), cela n'est cependant pas à exclure. Aussi, il existe plusieurs techniques pour détecter ce phénomène (e.g. test de *Klein*, Facteur d'inflation de la variance, etc.). Mais la plus simple consiste à calculer la matrice des corrélations entre les variables explicatives pour sélectionner ensuite celles qui ne sont pas fortement corrélées entre elles. Pratiquement, deux variables sont considérées comme colinéaires si leur coefficient de corrélation est supérieur à 0,8 ; dans ce cas, l'une d'entre elles doit être exclue du calcul du modèle de régression.

> **Dans le cas AFARC**, l'examen de la matrice des corrélations (tableau 6.15) montre que les variables NORMES, RISQUE, CROYANCES et EFFICACITÉ sont significativement corrélées entre elles, mais leurs coefficients de corrélation varient entre 0,411 et 0,631. Par ailleurs, elles sont corrélées avec la variable ATTITUDE qui est à expliquer : ce lien est indispensable pour espérer un effet significatif des variables explicatives sur la variable à expliquer. On

peut observer également que les variables AGE et REVENU ne sont corrélées significativement, ni avec les autres variables explicatives, ni avec la variable à expliquer, ce qui explique leur rejet du modèle de régression.

En amont des calculs nécessaires pour construire le modèle régressif, il est prudent d'analyser les corrélations entre les variables impliquées dans la régression pour éviter tout problème de colinéarité, ou de multicolinéarité quand cela concerne plusieurs variables à la fois, afin de sélectionner des variables peu liées.

Tableau 6.15 – Matrice des corrélations bivariées

	ATTITUDE	NORMES	RISQUE	CROYANCES	EFFICACITÉ	AGE	REVENU
Attitude	1						
Normes	0,620**	1					
Risque	*0,591***	0,530**	1				
Croyances	*0,631***	0,616**	0,590**	1			
Efficacité	*0,496***	0,411**	0,431**	0,433**	1		
Age	*-0,014*	0,037	-0,047	0,032	0,001	1	
Revenu	*0,029*	0,020	0,013	0,018	-0,022	0,032	1

** La corrélation est significative au niveau 0,01 (test bilatéral)

Il existe bien d'autres méthodes plus complexes pour détecter la colinéarité entre les variables explicatives. Par exemple, le logiciel SPSS propose deux statistiques nommées *Tolérance* et *VIF (variance-inflation factors)* qui permettent d'évaluer la proximité linéaire entre deux ou plusieurs variables explicatives.

Dans la pratique, étant donné qu'en science de gestion on rencontre rarement des variables totalement indépendantes, quand les estimations du modèle restent statistiquement significatives, on accepte souvent les résultats, même en présence d'un peu de colinéarité.

Focus

Deux indices d'évaluation rapide de la multicolinéarité

Les deux statistiques *Tolérance* et *VIF* sont bâties en calculant pour chaque variable explicative le coefficient de détermination multiple du modèle de régression linéaire sur les autres variables explicatives.

La Tolérance exprime la part de variance d'une variable explicative donnée qui n'est pas expliquée par les autres variables explicatives. Le niveau de *Tolérance* est calculé avec la formule :

$$\text{Tolérance} = 1 - R^2$$

☞

© Dunod – Toute reproduction non autorisée est un délit.

☞

Ce paramètre doit être le plus élevé possible. Généralement on accepte un niveau de tolérance allant jusqu'à 0,2. Pour des valeurs inférieures à 0,1, les prédicteurs doivent être manipulés avec précaution. Si la *Tolérance* est inférieure à 0,01, il est conseillé d'exclure la variable explicative concernée.

Le *Vif* n'est autre que l'inverse de la *Tolérance*. Il est calculé avec l'expression suivante :

$$Vif = \frac{1}{1 - R^2}$$

Le *Vif* utilise la même information que la *Tolérance* et s'interprète comme suit : idéalement, il faut obtenir une valeur de 1 (absence de corrélation entre les variables prédictives). Si une valeur de 1 à 5 est acceptable, à partir de 10 la variable devient suspecte et elle doit être exclue pour des valeurs supérieures à 100.

Application au cas AFARC

Calcul de ces deux statistiques pour le modèle avec quatre prédicteurs liés à la variable dépendante Attitude. En calculant le modèle de régression de la variable explicative Normes sur les trois autres variables explicatives (Efficacité, Risque et Croyances), le coefficient de détermination multiple vaut $R^2 \cong 0{,}436$, ce qui correspond à une *Tolérance* de 1-0,435 ≅ 0,565 et un *VIF* de 1/0,565 ≅ 1,773.

Le tableau 6.16 donne ces statistiques pour les six prédicteurs. On peut alors y observer que pour les variables Age et Revenu, ces deux statistiques sont proches de 1 : elles n'ont donc pas de problèmes de colinéarité, d'autant plus que leurs corrélations avec les autres prédicteurs ne sont pas du tout significatives.

Tableau 6.16 – Statistiques de colinéarité

Prédicteurs	Tolérance	VIF
Normes	0,564	1,773
Risque	0,578	1,729
Croyances	0,509	1,965
Efficacité	0,747	1,339
Âge	0,989	1,011
Revenu	0,997	1,003

Par ailleurs, la statistique de *Tolérance* de la variable Efficacité est assez élevée, de même que pour les trois autres variables : elle est toujours supérieure à 0,2 (et *VIF* < 5), ce qui signifie qu'aucune des variables de l'échantillon étudié n'a de sérieux problème de colinéarité.

7 La normalité et l'indépendance des résidus

Comme déjà précisé en début de chapitre, l'une des conditions d'application de la régression est la normalité des résidus. Il existe plusieurs méthodes pour le vérifier : tracer la distribution des résidus, tracer la droite de Henry, normalité par moitié,

écarts à la normalité, etc. La droite d'Henri, par exemple, est constituée du nuage de points mettant en relation les valeurs observées des résidus en abscisse et les valeurs théoriques de la distribution normale en ordonnées.

L'indépendance des résidus (e_i) constitue l'une des conditions les plus contraignantes de la régression. L'indépendance des données peut être observée visuellement à l'aide du nuage des points des résidus. Le test de Durbin-Watson est la technique par excellence pour vérifier l'indépendance des résidus et, par conséquent, celle des variables explicatives. Son principe repose sur l'étude de l'autocorrélation des résidus et consiste à calculer le rapport entre : la somme de la différence au carré des résidus entre les i et les $(i - 1)$ et la somme totale des résidus :

$$d = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2}$$

Par construction, cette statistique varie entre 0 et 4. Si d est proche de 0, il y a une autocorrélation positive entre les valeurs de la variable « résidus » ; mais si d est proche de 4, l'autocorrélation est négative. Si d varie autour 2, il n'y a pas d'autocorrélation (indépendance). Une interprétation plus précise (mais plus complexe) peut être obtenue à l'aide d'une table des valeurs critiques (notées *dl* et *du*) dépendante du nombre d'observations dont on dispose (habituellement moins de 15).

Remarque

Il est possible d'inclure des variables catégorielles comme prédicteurs dans des modèles de régression. Pour ce faire, il faut transformer auparavant les variables catégorielles en variables muettes (*i.e.* recodées 0/1), avec k-1 variables muettes si k est le nombre des modalités de la variable catégorielle). L'interprétation des coefficients de régression avec des variables muettes est différente de celle pour les variables métriques. Comme son nom l'indique, la régression linéaire multiple fait appel à un modèle linéaire simple de type $aX+bY+\ldots$, mais d'autres modèles plus complexes de type polynomial, exponentiel, puissance, etc. peuvent être utilisés. Le calcul de coefficients pour les modèles linéaire, polynomial, exponentiel et puissance s'effectue par la méthode des moindres carrés. Pour les modèles polynomial et exponentiel, il faut préalablement effectuer une transformation logarithmique pour les rendre linéaires.

Pour aller plus loin – Autres modèles de régression

Régression logistique

La régression logistique permet de prédire ou d'expliquer des variables binaires à partir d'une série de variables binaires d'origine ou rendues binaires (modèle binomial). Lorsque la variable à prédire possède plusieurs modalités non ordonnées, on parle de régression logistique multinomiale. La régression logistique est très utilisée en

☞

© Dunod – Toute reproduction non autorisée est un délit.

☞

marketing et en finance, dans le domaine des assurances et en médecine. Les fonctions les plus couramment utilisées en régression logistique sont la fonction logistique (modèle Logit), la fonction de répartition de la loi normale standard (modèle Probit), la fonction Log-log complémentaire et la fonction de Gompertz.

Régression Linéaire Généralisée (PLS)

La régression linéaire généralisée est un modèle non linéaire prédisant une variable dépendante à partir d'un ensemble de prédicteurs métriques et/ou non-métriques. Il permet d'unifier les autres modèles statistiques y compris la régression linéaire et la régression logistique qui deviennent des cas particuliers du modèle linéaire généralisé. On distingue le modèle PLS usuel qui utilise une méthode itérative basée sur des moindres carrés ordinaires (*Partial Least Squares*) et le modèle généralisé qui fait une utilisation itérative du maximum de vraisemblance (*Projection onto Latent Structure*).

L'ESSENTIEL

■ *Objectifs de la régression*

Les modèles de régression usuels sont des méthodes qui rendent compte de l'existence d'une relation linéaire entre variables métriques, où les variations de l'une – la variable dépendante – est statistiquement expliquée, soit par les variations d'une seule variable explicative : c'est la régression simple (RLS) ; soit par celles de plusieurs variables explicatives indépendantes : c'est la régression multiple (RLM). Quand le modèle établi est valide, cette relation peut être utilisée pour prévoir les variations d'une variable dépendante pour des valeurs précises des variables prédictives. Ces deux modèles sont en fait un cas particulier de la régression linéaire généralisée qui est un modèle non linéaire permettant de traiter des variables non métriques.

■ *Conditions d'application d'une régression*

La régression étant un modèle probabiliste et non déterministe, les coefficients de régression sont calculés pour ajuster au mieux les données estimées par le modèle aux données observées en minimisant les écarts de prévision. Un certain nombre d'opérations sont nécessaires pour valider la significativité de ces coefficients et la réalité de la relation linéaire. Dans une logique prédictive, il importe aussi de vérifier l'opérationnalité du modèle en s'assurant que les valeurs prédites sont suffisamment précises pour que le modèle régressif établi présente un réel intérêt managérial.

Section 3 LA PRÉVISION FONDÉE SUR DES SÉRIES TEMPORELLES

Beaucoup de données utilisées en gestion, comme en économie, varient en fonction du temps. C'est le cas, par exemple, du chiffre d'affaires d'une entreprise, du nombre de produits fabriqués par unité du temps, de l'évolution du prix d'une action, d'un indice boursier, ou du volume des sinistres traités par une compagnie d'assurance, etc. La variation d'une valeur quelconque en fonction du temps constitue une série temporelle, aussi appelée série chronologique (« *time series* » en anglais).

La particularité d'une telle série temporelle, qui correspond à la distribution conjointe de deux variables, est que l'une d'elles représente le temps ordonné. L'unité de temps, appelée aussi *période*, peut être exprimée en jours, semaines, trimestres, semestres, années, etc., avec la seule restriction d'être une mesure constante (notion de *périodicité*). La courbe qui décrit l'évolution d'une valeur en fonction du temps peut prendre diverses formes (linéaire, exponentielle, irrégulière, etc.) et, bien souvent, cette courbe dessine des fluctuations qui sont plus ou moins amples et répétitives au cours du temps.

1 Les composantes d'une série temporelle

Les séries temporelles, notées Y_t, sont le plus souvent le produit de plusieurs composantes, telles que la *tendance*, la *composante saisonnière* et la *composante résiduelle*, pour ne citer que les plus importantes. Il en existe d'autres, comme la *composante cyclique* par exemple, la composante due aux effets accidentels ou irréguliers, dont la détermination est difficile.

Définitions

La tendance, notée F_t. La tendance (ou *trend*) d'une série temporelle – appelée aussi composante fondamentale – reflète l'évolution générale à moyen ou long terme (croissante, décroissante ou stable) du phénomène étudié. Le *trend* intègre la totalité des effets permanents auxquels la série est soumise. Cette composante est estimée, soit avec des méthodes paramétriques, soit avec des méthodes de lissage.

La composante saisonnière, notée S_t, décrit des cycles plus ou moins courts et répétitifs dans le temps, sur une période p donnée qui doit être constante (*e.g.* une année) : c'est le cas, par exemple, du chiffre d'affaires d'une entreprise de distribution de boissons rafraîchissantes dont la fluctuation résulte bien des saisons). Cette composante est donc une variation saisonnière qui se reproduit avec régularité dans le temps. Considérée comme strictement périodique ($S_{t+p} = S_t$), elle permet de distinguer, à l'intérieur d'une

© Dunod – Toute reproduction non autorisée est un délit.

période, les effets positifs ou négatifs sur la valeur mesurée. La composante saisonnière est définie par les coefficients saisonniers (S_j, où j=1, 2, ..., p) de sorte que $\sum S_j = 0$.

La composante résiduelle, notée E_t, traduit les effets non structurés du phénomène étudié (appelé aussi « bruit blanc gaussien »). Elle est modélisée par une suite de variables aléatoires indépendantes, centrées, et de variance identique.

La composante cyclique : Elle rend compte d'effets dont la longueur de la période est trop importante (très long terme) pour être qualifiée de moyen ou long terme ; c'est le cas de l'effet des crises économiques ou de cycles d'activité. Comme il est difficile en pratique de la dissocier des autres effets périodiques, elle est par conséquent prise en compte dans la tendance F_t.

L'analyse des séries temporelles peut être perturbée par des facteurs qu'il est difficile de gérer. Dans ces conditions, on essaie de réduire ou d'éliminer leur effet en les retirant des analyses :

Les effets accidentels (ou anormaux). Ce sont des effets qui modifient très sensiblement l'évolution de la série temporelle. Comme l'indique leur nom, ils sont de nature imprévisible ou aléatoire (*e.g.* pluviométrie très élevée, catastrophe naturelle) ou accidentelle (nombre de voitures sur une route à cause d'un accident). Comme il est quasiment impossible de prévoir ces événements anormaux, il est d'usage de « nettoyer » les séries temporelles étudiées avant leur analyse, en supprimant ces effets accidentels.

Perturbation des jours ouvrés. Beaucoup de données, d'ordre économique ou financier, dépendent des jours ouvrés (à l'exemple du cours des actions, la bourse restant fermée en fin de semaine) : il est donc d'usage d'exclure le week-end et les jours fériés des séries étudiées.

L'intérêt pour une entreprise d'analyser des séries de données en fonction du temps est de pouvoir faire des prévisions sur ce que le futur pourrait être à partir de l'historique des données du passé. Il faut cependant pouvoir se fier à ces prévisions. Pour ce faire, il est nécessaire de simplifier les séries temporelles brutes (par lissage par exemple) ou de les décomposer (pour identifier les variations saisonnières) afin de rechercher des modèles théoriques qui décrivent le plus fidèlement possible l'évolution temporelle des valeurs avec l'objectif de faire des prévisions fiables, au moins pour le court terme.

Cette section traite d'abord de la préparation des données brutes pour rendre leur analyse possible ; puis des méthodes de lissage utilisées pour identifier la tendance générale d'une série temporelle, avant de présenter des méthodes pour analyser la composante saisonnière.

L'exemple concernant l'évolution du chiffre d'affaires de l'entreprise DOREST sur cinq années consécutives est utilisé pour illustrer l'application de ces méthodes. Le tableau 6.17 donne l'évolution du CA trimestriel exprimé en milliers d'euros.

Cas d'illustration

DOREST en progression permanente depuis 2010

Localisée au sud des États-Unis, la firme DOREST, entreprise produisant des boissons rafraîchissantes depuis plus d'un siècle, connaît depuis 2010 une progression constante de son chiffre d'affaires. Le tableau 6.17 témoigne de cette évolution mensuelle entre 2010 – période à laquelle l'entreprise a commencé son ascension suite à quelques années difficiles en raison de la crise de 2008 – et 2015. Le PDG de l'entreprise aimerait savoir comment pourrait évoluer son chiffre d'affaires si, dans les prochains mois, le développement de son activité continuait sur le rythme actuel (« *business as usual* »), afin d'envisager les investissements nécessaires qui pourraient, en conséquence, s'imposer.

Tableau 6.17 – Évolution du chiffre d'affaires de l'entreprise DOREST pour la période 2010-2015

Période	Chiffre d'affaires, milliers d'euros					
	2010	2011	2012	2013	2014	2015
JANVIER	6817	6737	7602	8372	8175	8750
FEVRIER	8118	8255	9051	8973	9622	10254
MARS	8095	9245	9495	9765	10540	10865
AVRIL	9076	9816	10412	10096	11193	11751
MAI	8349	9267	9998	10790	10423	11224
JUIN	7959	8072	8710	9615	9411	10542
JUILLET	6550	7874	7461	8166	8544	9651
AOUT	6292	6165	6932	7365	7626	8036
SEPTEMBRE	4937	5806	6243	6560	7417	7694
OCTOBRE	4977	5546	5978	6152	6821	7505
NOVEMBRE	5757	5578	6589	6529	7882	7681
DECEMBRE	5816	6576	7447	7354	7681	8562

2 Préparer les données brutes

Avant de se lancer dans l'analyse d'une série temporelle, il est indispensable d'effectuer un certain nombre de vérifications afin d'éviter de compromettre la validité de l'analyse.

© Dunod – Toute reproduction non autorisée est un délit.

2.1 La vérification de la qualité des données

Ces vérifications préalables concernent :

– la régularité des mesures (ex. : le nombre de jours de production selon les mois) ;
– la stabilité des séries : constance de la fréquence et de l'amplitude des effets périodiques ;
– la permanence de la mesure du paramètre étudié sur toute la durée analysée (par exemple, la modification du mode de calcul du taux de chômage a modifié les valeurs calculées).
– l'existence d'une réelle variation périodique (la composante saisonnière), faute de quoi les techniques d'analyse des variations saisonnières ne pourront pas être utilisées.

D'où une triple tâche à effectuer pour préparer les données afin qu'elles soient exploitables :

– **La réorganisation des données :** elles se présentent souvent sous la forme d'un tableau à double entrée, à l'exemple du tableau 6.17. Les données doivent être réorganisées pour obtenir un tableau avec deux colonnes dont la première contient les périodes de temps dans l'ordre croissant.

> Dans l'exemple DOREST, la première colonne des données correspond au nombre croissant des mois de 1 à 72 comme le montre la figure 6.18 : la valeur 65 correspond au mois de mai de l'année 2015).

– **La correction des jours ouvrables.** Certaines données – par exemple les données économiques et financières – sont très sensibles au nombre de jours ouvrables. Lorsque les périodes sont des semaines, des mois, voire des années, il convient alors d'effectuer une correction sur le nombre de jours ouvrables inclus dans la période définie.

> Par exemple, la série temporelle des valeurs boursières du CAC40 ne contient pas les week-ends et les jours fériés.

– **La correction des valeurs anormales**. Il convient encore de corriger les valeurs irrégulières ou accidentelles issues d'aléas imprévisibles.

2.2 Les principes de notation utilisés

– Les valeurs observées dans une série temporelle liée à une période t donnée sont notées y_t ;
– les valeurs prédites de la variable Y pour une période donnée sont notées $\hat{y}_t$. Il est donc possible de mettre en relation les valeurs prédites avec les valeurs observées pour n'importe quelle période t, mais il faut entendre par « prévision », les valeurs de Y prédites dans le futur proche (soit, de manière usuelle : $\hat{y}_{t+1}$, $\hat{y}_{t+2}$, …).

– La précision d'une prévision est évaluée avec l'*erreur de prévision* définie comme étant le rapport de la somme du carré des écarts entre les valeurs prédites et les valeurs observées, sur le nombre de valeurs comprises dans la série temporelle.

La suite de cette section est focalisée :

– d'abord sur l'analyse du *trend* (la tendance générale), analyse dont l'objectif est d'isoler cette composante de l'ensemble des autres avec l'objectif de faire les prévisions de base ;
– puis, sur l'analyse de la composante saisonnière pour vérifier la stabilité des effets saisonniers et obtenir, en conséquence, la série temporelle corrigée de variations saisonnières (SCV) qui, au même titre que les séries lissées, peut être utilisée à son tour pour faire des prévisions.

3 Analyser la tendance

L'analyse de la tendance générale (qui inclut souvent la composante cyclique) peut s'effectuer par l'ajustement d'un modèle théorique, tel que le modèle linéaire si les valeurs de la série s'alignent approximativement autour d'une droite ou d'autres modèles plus complexes tels que les modèles polynomial, exponentiel, puissance, logistique, etc. Cette partie ne traite que de deux techniques simples d'analyse de la tendance : l'ajustement linéaire et les moyennes mobiles d'ordre p.

3.1 L'analyse de la tendance par l'ajustement linéaire

Cette méthode consiste à calculer les coefficients de la droite des moindres carrés $y = b_0 + b_1 t$ qui décrit le nuage des points composant la série, selon la procédure décrite pour la régression linéaire (RLS), où la variable explicative est le temps (T).

$$b_1 = \frac{Cov(Y,T)}{Var(T)}$$

et

$$b_0 = \overline{Y} - b_1 \overline{T}$$

Dans le cas DOREST, avec la série des données brutes, on obtient $b_0 = 6991$ et $b_1 = 31$. Le modèle linaire est statistiquement significatif, mais ne rend compte que d'environ 15 % de la variance, ce qui s'explique par un effet assez prononcé de la composante saisonnière.

Le modèle linéaire rend bien compte de la tendance générale, mais il globalise aussi la saisonnalité des données. Pour avoir une vision juste de la tendance, il faut alors calculer la droite de régression sur la série nettoyée des variations saisonnières. Pour ce faire, on a recours soit aux moyennes mobiles centrées d'ordre p (voir section suivante), soit à la série des points qui représentent la moyenne des p périodes.

© Dunod – Toute reproduction non autorisée est un délit.

En utilisant la moyenne mobile d'ordre 12 (parce que la période couvre 12 mois), l'équation de la droite des moindres carrés, pour le cas DOREST, s'écrit y = 39,201*Période (en mois) + 6693, avec un R^2 = 0,9897. Le modèle linéaire explique alors presque la totalité de la variabilité temporelle du chiffre d'affaires. En utilisant les moyennes annuelles, l'équation devient y = 244,56*Période (en année) - 483452 ; avec un R^2 = 0,9974. La tendance est bien en augmentation, à raison de 39201euros par mois.

Il ne faut cependant pas oublier que la droite des moindres carrés propose une représentation abstraite qui ne correspond pas toujours fidèlement la réalité, laquelle peut être très complexe. Comme dit déjà pour la régression simple (*cf.* 5.3), il faut bien examiner l'intervalle de confiance des prévisions moyennes et individuelles pour évaluer la précision des résultats et pouvoir prendre une décision adaptée.

Remarque

La méthode qui consiste à remplacer observations par leur moyenne (laquelle représente alors la moyenne de p périodes) correspond à une technique de lissage appelée « moyennes échelonnées d'ordre p ». Par exemple, si l'évolution mensuelle des ventes en volume est connue sur n années, les moyennes échelonnées d'ordre 12 (ME_{12}) correspondent aux volumes annuels moyens des ventes. Ainsi, passe-t-on de n* 12 valeurs initiales à seulement n valeurs. Mais cette technique entraîne une perte d'information et ne permet plus d'analyser les fluctuations à l'intérieur des p périodes.

Dans l'exemple DOREST, le lissage de la série (dont les informations initiales sont données par mois) avec la méthode des moyennes échelonnées d'ordre p = 12, permet d'obtenir l'évolution annuelle du chiffre d'affaires. Avec cette méthode, il est important de bien situer le point de mesure moyen annuel sur l'axe du temps : ainsi, pour DOREST, le chiffre d'affaires annuel moyen considéré est situé au milieu de l'année, soit au 30 juin.

3.2 L'analyse de la tendance par les moyennes mobiles centrées d'ordre *p*

Cette technique de lissage est l'une des plus utilisées. Si elle ne permet pas de faire directement des prévisions (en raison d'une perte d'information qu'elle entraîne sur les données – comme il sera expliqué plus loin), elle révèle cependant très clairement le sens de l'évolution, la tendance de fond, de la série des valeurs étudiées.

Focus

Calcul de moyennes mobiles centrées

Soit une série temporelle qui contient N valeurs. L'ordre p de la moyenne mobile est souvent égal au nombre de mesures disponibles sur la période (ex. $p = 12$ si les données sont mensuelles, $p = 4$ si les données sont trimestrielles, etc.). La procédure de calcul de la moyenne mobile centrée diffère selon que l'ordre est un nombre pair ou impair.

Si p est un *nombre impair* (il existe une valeur centrale), on commence par calculer la moyenne de p premières valeurs selon la formule :

$$MM_p^1 = \frac{y_1 + y_2 + \ldots + y_p}{p}$$

Cette première valeur est alors placée en face de la valeur centrale des p premières valeurs. La valeur suivante est calculée avec la formule

$$MM_p^2 = \frac{y_2 + y_3 + \ldots + y_{p+1}}{p}$$

et ainsi de suite.

Si p est un *nombre pair*, il n'y a pas de valeur centrale : la moyenne de $p + 1$ premières valeurs est alors calculée selon la formule :

$$MM_p^1 = \frac{\frac{y_1}{2} + y_2 + \ldots + y_p + \frac{y_p}{2}}{p}$$

et on positionne la première moyenne mobile en face de la valeur centrale des $p + 1$ premières valeurs.

Les moyennes mobiles peuvent être influencées par les valeurs aberrantes et/ou extrêmes. Si la série analysée en contient, cet inconvénient est atténué en utilisant la médiane mobile à la place de la moyenne.

Comme les moyennes mobiles centrées se succèdent l'une derrière l'autre, il n'est donc pas possible de calculer la moyenne des premières et dernières valeurs de la série (n/2 si p est pair ou (n-1)/2 si p est impair). Cette perte d'information est un inconvénient majeur de l'usage des moyennes mobiles symétriques, car on ne dispose d'aucune information sur la tendance en début et en fin de période : en conséquence, elles ne permettent donc pas de faire des prévisions

Exemple, pour le cas DOREST, de calcul des moyennes mobiles symétriques d'ordre 12. C'est le mode de calcul des périodes paires qui est appliqué, puisque $p = 12$.

$$MM_{12}^1 = \frac{\frac{6817}{2} + 8118 + \ldots + 5816 + \frac{6737}{2}}{12} = 6892$$

$$MM_{12}^2 = \frac{\frac{8118}{2} + 8095 + \ldots + 6737 + \frac{8255}{2}}{12} = 6894$$

et ainsi de suite. Le tableau 6.18 donne les moyennes mobiles centrées d'ordre 6 et 12. Comme indiqué, les premières et les dernières valeurs de la série ne peuvent être calculées. Il en résulte que la courbe des moyennes mobiles d'ordre 12 est pratiquement une

© Dunod – Toute reproduction non autorisée est un délit.

droite qui fait clairement ressortir la tendance en hausse, alors que la courbe des moyennes mobiles d'ordre 6 contient encore des fluctuations bien qu'elle soit moins « accidentée » que la série de données initiale. C'est pour cette raison que, pour obtenir une tendance lisse dans le cas de séries présentant des variations saisonnières, l'ordre des moyennes mobiles doit correspondre aux périodes du cycle des variations saisonnières (12 dans le cas DOREST).

Tableau 6.18 – Exemple de calcul des moyennes mobiles centrées d'ordre 6 et 12 pour le cas DOREST

ANNEE	MOIS	t	CA	MM_6	MM_{12}
2010	JANVIER	1	6817	-	-
2010	FÉVRIER	2	8118	-	-
2010	MARS	3	8095	-	-
2010	AVRIL	4	9076	8047	-
2010	MAI	5	8349	7872	-
2010	JUIN	6	7959	7457	-
2010	JUILLET	7	6550	6852	6892
2010	AOÛT	8	6292	6295	6894
2010	SEPTEMBRE	9	4937	5900	6948
…	…	…	…	…	…
2015	MAI	65	11224	10530	9311
2015	JUIN	66	10542	10081	9340
2015	JUILLET	67	9651	9463	-
2015	AOÛT	68	8036	8813	-
2015	SEPTEMBRE	69	7694	8353	-
2015	OCTOBRE	70	7505	-	-
2015	NOVEMBRE	71	7681	-	-
2015	DÉCEMBRE	72	8562	-	-

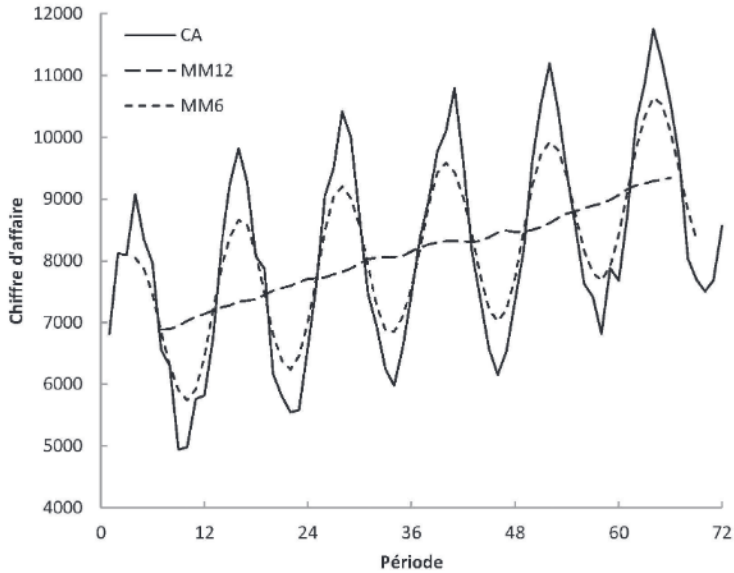


Figure 6.6 – Moyennes mobiles symétriques d'ordre 12 pour le cas DOREST

Si les moyennes mobiles centrées ne permettent pas de faire des prévisions, elles servent cependant à filtrer une série temporelle dans l'objectif de faire apparaître la tendance des valeurs étudiées. Pour éviter le problème des pertes de données, il est courant de calculer des moyennes mobiles non centrées (appelées parfois non symétriques, unilatérales, etc.) Si l'objectif de l'analyse est de prévoir la valeur à $\hat{y}_{t+1}$, cette dernière égalera la moyenne arithmétique simple des p dernières valeurs d'Y.

Remarque

Il existe bien d'autres types de moyennes mobiles : les moyennes mobiles « pondérées », « variables », « ajustées par les volumes », « triangulaires », etc., mais leur développement sort du cadre de cette section.

4 Détecter des variations saisonnières

Outre les représentations graphiques, la méthode la plus simple pour détecter la présence de variations saisonnières dans une série temporelle est le *coefficient d'autocorrélation*. Il s'apparente au coefficient de corrélation linéaire de Pearson établi entre deux ensembles de données dans une même série, mais décalés de k périodes :

– pour k=1 les séries sont : $y_1, y_2, \ldots y_{n-1}$ et $y_2, y_3, \ldots y_n$;
– pour k=2 les séries sont : $y_1, y_2, \ldots y_{n-2}$ et $y_3, y_4, \ldots y_n$; et ainsi de suite.

© Dunod – Toute reproduction non autorisée est un délit.

La première série correspond aux n-k premières valeurs et la deuxième correspond au n-k dernières valeurs de la même série initiale. Les deux séries ont toutes les deux le même nombre de termes (n-k) et pour des raisons liées au calcul, elles peuvent être considérées comme deux variables appariées. Pour chacune des deux séries, on peut alors calculer leur moyenne arithmétique $\bar{y}_{(1)}$ et $\bar{y}_{(2)}$, la covariance $Cov(y_{(1)},y_{(2)})$ et leurs écarts-type $ET(y_{(1)})$ et $ET(y_{(2)})$. Le coefficient d'autocorrélation d'ordre k est alors formulé comme suit :

$$r_k = \frac{Cov(y_{(1)}, y_{(2)})}{ET(y_{(1)})ET(y_{(2)})}$$

Comme le coefficient de corrélation, le coefficient d'autocorrélation varie entre -1 et +1. La graphique qui représente r_k en fonction de k est appelé *corrélogramme* (*cf.* figure 6.7). La présence des variations saisonnières dans une série temporelle se manifeste sous la forme de pics dans le corrélogramme en $k = p$, $k = 2p$, etc.

Le corrélogramme correspondant aux données du cas DOREST est présenté par la figure 6.7. Cette série temporelle représentant l'évolution du chiffre d'affaires contient bien une composante saisonnière dont la période p est de 12 mois.

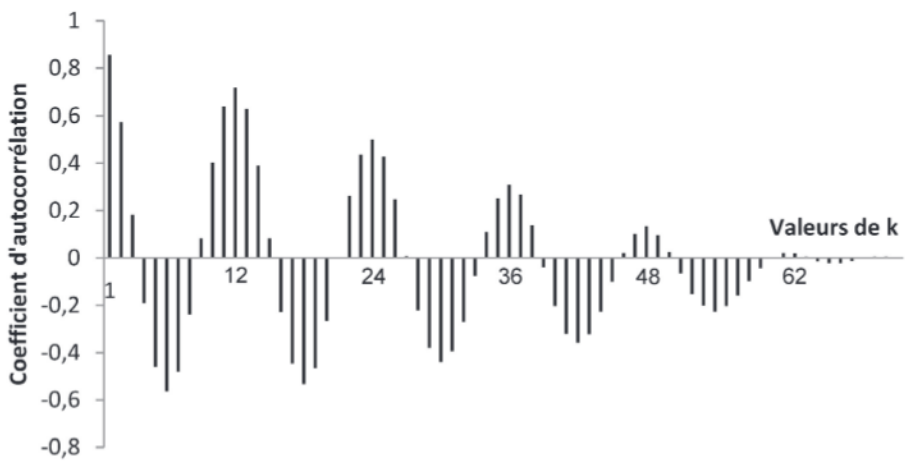


Figure 6.7 – Corrélogramme pour le cas DOREST

Dans l'analyse d'une série temporelle, il est important de pouvoir déceler l'existence d'une saisonnalité, à la fois pour des raisons techniques et opérationnelles. En effet, certaines méthodes de prévision, comme le lissage exponentiel simple par exemple, ne sont guère applicables sur des séries soumises à des variations saisonnières. D'autre part, la détermination du *trend* d'une série temporelle est beaucoup plus fiable si elle ne subit pas de variation saisonnière.

5 Désaisonnaliser les séries temporelles

Trois modèles différents permettent de décomposer une série temporelle contenant des variations saisonnières : les modèles : additif, multiplicatif et mixte. Pour choisir le modèle de décomposition le mieux adapté aux caractéristiques de la série, il existe plusieurs techniques graphiques et analytiques (dont la description détaillée dépasse cependant l'objectif de cet ouvrage).

5.1 Le modèle additif

Si les variations saisonnières s'ajoutent simplement à la tendance, c'est le modèle additif qui s'applique. Ce modèle considère la série temporelle Y_t comme la somme de la tendance générale F_t et des variations saisonnières S_t, donc : $Y_t = F_t + S_t$.

Selon ce modèle, la composante saisonnière S_t est obtenue en retranchant la tendance générale F_t et les variations résiduelles à la série brute Y_t : donc $S_t = Y_t - F_t$. La tendance peut alors être calculée avec l'une des méthodes déjà développées dans cette section (moyennes mobiles, droite d'ajustement, etc.). Si les moyennes mobiles sont utilisées pour déterminer la composante saisonnière, elles doivent être centrées et d'ordre fini.

> Calcul pour le cas DOREST de la série temporelle du chiffre d'affaires corrigée des variations saisonnières. Le résultat de ces calculs est présenté dans le tableau 6.19. La tendance générale F_t est estimée à l'aide des moyennes mobiles centrées d'ordre 12 (MM_{12}), car les données initiales sont disponibles par mois.
>
> Après le calcul des moyennes mobiles (MM_{12}), les écarts saisonniers sont calculés en soustrayant la tendance générale à la série brute ($S_t = Y_t - MM_{12}$). Ainsi, les écarts saisonniers sont obtenus pour chacun des 72 mois et ils expriment la part attribuée aux variations saisonnières. On peut observer que ces écarts diffèrent légèrement : par exemple, pour le troisième mois (mars), les écarts saisonniers sont de : 1966 (en 2011), 1713 (en 2012), 1469 (en 2013), 1994 (en 2014), 1614 (en 2015).
>
> Il faut rappeler que, pour chacun des 12 mois, il n'y a que 5 écarts saisonniers sur les 6 possibles, en raison de la méthode des moyennes mobiles centrées utilisée pour le calcul. Ce problème ne se rencontrerait pas si le calcul de la tendance générale avait été opéré avec la méthode de la droite d'ajustement.

L'étape suivante consiste à calculer les coefficients saisonniers $\bar{S}_t$. Ces coefficients sont obtenus en calculant mois par mois la moyenne arithmétique des écarts entre les valeurs de la série brute (CA) et celles de la série lissée (MM_{12} dans l'exemple). Ainsi, dans l'exemple, le coefficient saisonnier pour le mois de janvier est calculé avec la formule :

$$\bar{S}_1 = \frac{(-462) + (-115) + \ldots + (-409)}{5} = -226$$

Les 11 autres coefficients saisonniers sont calculés de la même manière et sont ensuite placés en face de chaque valeur brute trimestrielle. Finalement, la série tem-

© Dunod – Toute reproduction non autorisée est un délit.

porelle corrigée des variations saisonnières, appelée CVS est bien obtenue en retranchant les coefficients saisonniers à la série brute. Le graphique de la figure 6.8 représente la série brute et la même série corrigée des variations saisonnières telle que calculée avec le modèle additif. Ainsi, la série lissée qui est obtenue peut être utilisée pour effectuer des prévisions à court terme.

Tableau 6.19 – Décomposition du chiffre d'affaires – modèle additif

t	CA	MM_{12}	$S_t = Y_t - MM_{12}$	$\bar{S}_t$	CVS
1	6817		-	-226	7043
2	8118		-	1037	7081
3	8095		-	1751	6344
4	9076		-	2378	6698
5	8349		-	2028	6321
6	7959		-	919	7040
7	6550	6892	-342	-175	6725
8	6292	6894	-602	-1052	7344
9	4937	6948	-2011	-1776	6713
10	4977	7027	-2050	-2119	7096
11	5757	7096	-1339	-1593	7350
12	5816	7139	-1323	-1131	6947
13	6737	7199	-462	-226	6963
…	…		…	…	…
65	11224	9311	1913	2028	9 196
66	10542	9340	1202	919	9623
67	9651		-	-175	9826
68	8036		-	-1052	9088
69	7694		-	-1776	9470
70	7505		-	-2119	9624
71	7681		-	-1593	9 274
72	8562		-	-1131	9 693

Coefficients saisonniers	
$\bar{S}_1$	-226
$\bar{S}_2$	1037
$\bar{S}_3$	1751
$\bar{S}_4$	2378
$\bar{S}_5$	2028
$\bar{S}_6$	919
$\bar{S}_7$	-175
$\bar{S}_8$	-1052
$\bar{S}_9$	-1776
$\bar{S}_{10}$	-2119
$\bar{S}_{11}$	-1593
$\bar{S}_{12}$	-1131

Comment prédire le chiffre d'affaires de l'entreprise DOREST en janvier 2016 à partir du modèle additif ? Pour obtenir le chiffre d'affaires annuel moyen, l'une des solutions serait d'utiliser la droite des moindres carrés pour déterminer d'abord la tendance générale et l'ajuster ensuite avec les variations saisonnières. L'équation de la droite d'ajustement est $F_t = 6664 + 40 * Période$. La valeur de la tendance en janvier 2016 est alors $F_{73} = 6664 + 40 * 73 = 9584$ et donc le chiffre d'affaires prévisionnel serait

$$\hat{y}_t = F_t + S_t = 9584 - 226 = 9358$$

L'application de cette formule aux autres mois de 2016 fournit les 12 prévisions mensuelles du chiffre d'affaires de 2016 :

Janvier	Février	Mars	Avril	Mai	Juin
9358	10661	11415	12082	11772	10703
Juillet	Août	Septembre	Octobre	Novembre	Décembre
9649	8812	8128	7825	8391	8893

La figure 6.8 montre l'évolution prévisionnelle du chiffre d'affaires en 2016 (période 73-84 mois). Mais cette prévision doit être interprétée avec beaucoup de précaution, surtout pour les valeurs éloignées des dernières données de l'historique, car elles ne seront valables que si les conditions de l'activité de l'entreprise restent exactement les mêmes qu'auparavant.

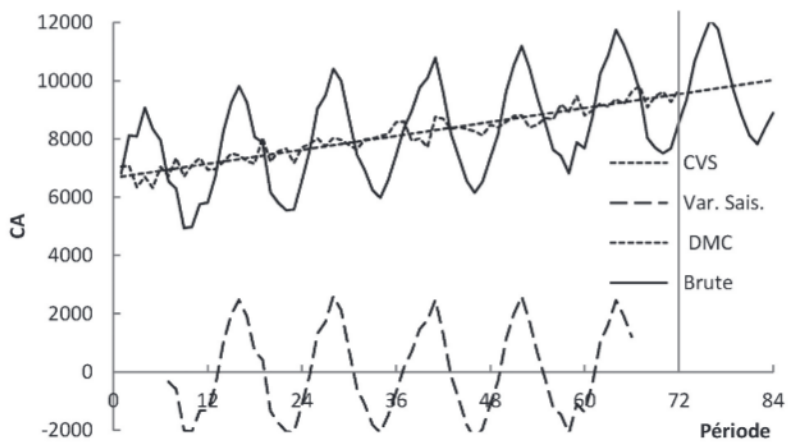


Figure 6.8 – Série corrigée des variations saisonnières et prévision – modèle additif

5.2 Le modèle multiplicatif

Le modèle multiplicatif s'applique quand les variations saisonnières sont proportionnelles à la tendance. Dans ce cas, la série temporelle peut être considérée comme étant le produit entre la tendance générale (F_t) et les variations saisonnières (S_t) : $Y_t = F_t \times S_t$. Pour déterminer la composante saisonnière (S_t), il faut donc diviser la série brute par la tendance : $S_t = Y_t/F_t$.

La démarche pour évaluer la composante saisonnière et calculer la série corrigée des variations saisonnières (CVS) avec le modèle multiplicatif est très similaire à celle du modèle additif. La principale différence concerne le calcul des écarts saisonniers (S_t) et de leurs moyennes $\bar{S}_t$. Étant donné qu'il s'agit de rapports, les coefficients saisonniers S_t correspondent à la moyenne géométrique des écarts saisonniers pour chaque période initiale. Le résultat des calculs est présenté dans le tableau 6.20. La dernière

© Dunod – Toute reproduction non autorisée est un délit.

colonne du tableau donne les valeurs calculées en utilisant la droite des moindre carrée (DMC) sur la série CVS. Comme dans le cas du modèle additif, la tendance générale est estimée à l'aide des moyennes mobiles centrées d'ordre 12. Ainsi, le coefficient saisonnier pour le mois de janvier est calculé avec la formule :

$$\bar{S}_1 = \sqrt[5]{0{,}936 * 0{,}985 * 1{,}019 * \ldots * 0{,}955} = 0{,}972$$

Tableau 6.20 – Décomposition de la série brute - modèle multiplicatif

t	Y_t	MM_{12}	$S_t = Y_t - MM_{12}$	$\bar{S}_t$	CVS	DMC
1	6817	-	-	0,972	7017	6704
2	8118	-	-	1,127	7200	6744
3	8095	-	-	1,214	6666	6784
4	9076	-	-	1,289	7043	6824
5	8349	-	-	1,245	6704	6864
6	7959	-	-	1,109	7176	6904
7	6550	6892	0,950	0,977	6702	6944
8	6292	6894	0,913	0,867	7253	6984
9	4937	6948	0,711	0,773	6387	7024
10	4977	7027	0,708	0,734	6780	7064
11	5757	7096	0,811	0,799	7206	7104
12	5816	7139	0,815	0,859	6772	7144
13	6737	7199	0,936	0,972	6934	7184
14	8255	7248	1,139	1,127	7322	7224
…	…	…	…	…	…	…
65	11224	9311	1,205	1,245	9013	9264
66	10542	9340	1,129	1,109	9505	9304

Coefficients saisonniers	
$\bar{S}_1$	0,972
$\bar{S}_2$	1,127
$\bar{S}_3$	1,214
$\bar{S}_4$	1,289
$\bar{S}_5$	1,245
$\bar{S}_6$	1,109
$\bar{S}_7$	0,977
$\bar{S}_8$	0,867
$\bar{S}_9$	0,773
$\bar{S}_{10}$	0,734
$\bar{S}_{11}$	0,799
$\bar{S}_{12}$	0,859

La série temporelle corrigée des variations saisonnières (CVS) est alors obtenue *en divisant* la série temporelle brute par les coefficients saisonniers moyens de chaque période. La figure 6.9 représente graphiquement les différentes composantes calculées et les prévisions pour 2016. On peut remarquer que, quel que soit le modèle utilisé – additif ou multiplicatif –, les deux séries temporelles obtenues et corrigées des variations saisonnières sont très voisines. Tout comme la tendance, la série temporelle corrigée peut être utilisée pour effectuer des prévisions à court terme.

> Dans le cas DOREST, la valeur de la tendance en janvier 2016 valant , le chiffre d'affaires prévisionnel serait donc :
>
> $$\hat{y}_t = F_t * S_t = 9584 * 0{,}972 = 9316$$

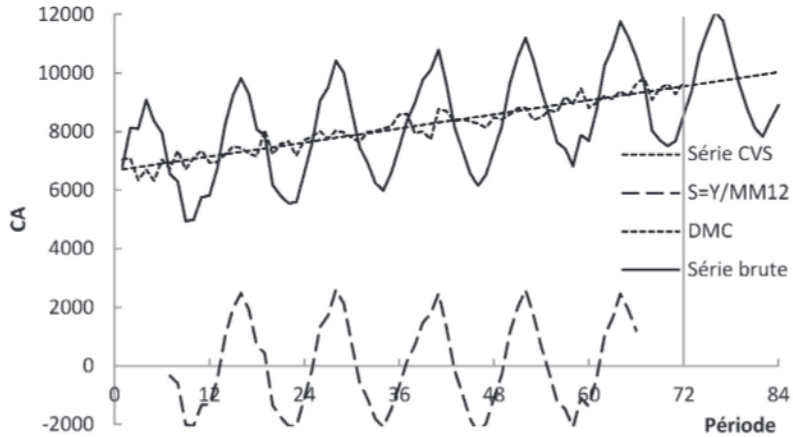


Figure 6.9 – Décomposition des données brute et prévision – modèle multiplicatif

5.3 Le modèle mixte

Le modèle mixte, combinant additif et multiplicatif, est utilisé quand les variations saisonnières sont proportionnelles à la tendance (multiplicatif) alors que les fluctuations irrégulières saisonnières s'ajoutent simplement à la tendance (additif).

6 Utiliser les méthodes de prévisions par lissage

Il existe un grand nombre de techniques quantitatives et qualitatives pour prévoir des séries temporelles. Les techniques quantitatives sont presque exclusivement basées sur les séries de données des valeurs passées (historique) alors que les techniques qualitatives, effectuées par des experts, sont plutôt utilisées lorsque les données historiques sont insuffisantes, voire indisponibles, pour effectuer des prévisions. Seules les techniques quantitatives empiriques sont développées ici.

Les techniques quantitatives de prévision sont réputées fournir un niveau de précision élevé et, par conséquent, apparaissent comme une aide indispensable à la prise de décision. Fondées sur l'exploitation des données historiques, les moyennes mobiles simples et doubles, le lissage exponentiel simple et double, les méthodes de Holt et de Holt-Winters sont parmi les techniques quantitatives les plus utilisées.

6.1 La prévision par les moyennes mobiles non centrées

Le recours aux moyennes mobiles non centrées permet de pallier les inconvénients de l'usage des moyennes mobiles centrées (*cf.* § 3.2 : perte des données en début et

© Dunod – Toute reproduction non autorisée est un délit.

fin de série empêchant d'effectuer une bonne prévision). Cela consiste à prédire la valeur $\hat{y}_{t+1}$ en calculant la moyenne arithmétique des p dernières valeurs d'Y.

La figure 6.10 montre le graphique des cours du CAC 40 avec les moyennes mobiles non centrées d'ordre 20, 50 et 100 telles qu'affichées sur le site www.zone-bourse.com. Il faut noter que ce graphique ne contient pas de valeurs manquantes en fin de série parce qu'il repose sur les moyennes mobiles non centrées. Les moyennes mobiles de la période t sont calculées en faisant la moyenne des 20, 50 ou 100 dernières observations avant cette période.



Figure 6.10 – Moyennes mobiles non centrées d'ordre 20, 50 et 100 – CAC40

Soit la prévision de l'évolution du chiffre d'affaires de DOREST en 2015 à court et à moyen terme. Le tableau 6.21 illustre le calcul des moyennes mobiles non centrées d'ordre 2, 3 et 4. Ainsi la prévision du chiffre d'affaires pour le mois de janvier 2016 serait, en utilisant les MM_2 : (9 274+9 693)/2=9484, les MM_3 : 9 530 et les MM_4 : 9 515.

Mais comment savoir laquelle de ces trois prévisions est la meilleure ? Pour ce faire, il faut calculer la somme des carrés des écarts entre la série brute et la prévision (comme le montrent les trois dernières colonnes du tableau). On peut observer que les erreurs les plus faibles (725 530) sont obtenues avec les MM3. C'est donc les MM3 qui fournissent la meilleure prévision. La figure 6.11 compare graphiquement les quatre séries (brute, puis avec MM2, MM3 et MM4).

Tableau 6.21 – L'évolution du CA (CSV) de DOREST en 2015

Périodes	Série brute	MM2	MM3	MM4	Erreurs MM2	Erreurs MM3	Erreurs MM4
1	8 976	-	-	-	-	-	-
2	9 217	-	-	-	-	-	-
3	9 114	9 097	-	-	289	-	-
4	9 373	9 166	9 102	-	42 849	73 441	-
5	9 196	9 244	9 235	9 170	2 304	1 521	676
6	9 623	9 285	9 228	9 225	114 244	156 025	158 404
7	9 826	9 410	9 397	9 327	173 056	184 041	249 001
8	9 088	9 725	9 548	9 505	405 769	211 600	173 889
9	9 470	9 457	9 512	9 433	169	1 764	1 369
10	9 624	9 279	9 461	9 502	119 025	26 569	14 884
11	9 274	9 547	9 394	9 502	74 529	14 400	51 984
12	9 693	9 449	9 456	9 364	59 536	56 169	108 241
13	-	**9 484**	**9 530**	**9 515**	*991 770*	*725 530*	*758 448*

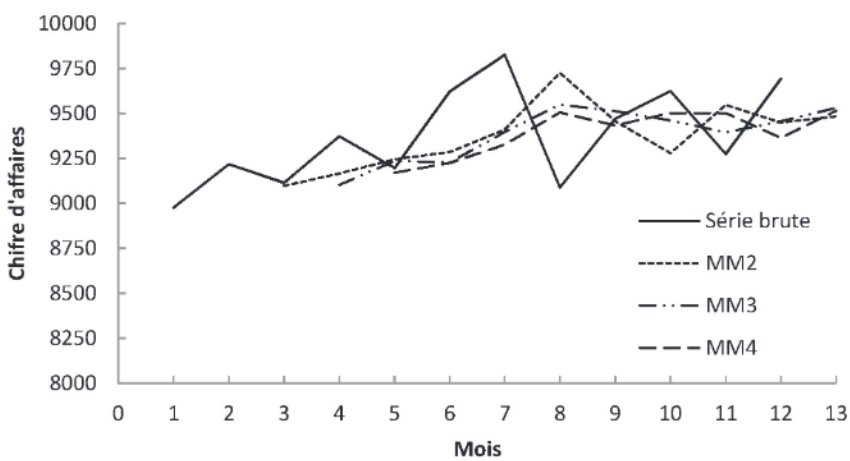


Figure 6.11 – Moyennes mobiles non centrées – Cas DOREST

Malgré leurs avantages (par comparaison aux moyennes mobiles centrées), les moyennes mobiles non centrées présentent trois importantes limites :

– un grand nombre de valeurs sont nécessaires pour effectuer la prévision ;
– toutes les p dernières observations entrant dans le calcul ont un poids égal ;
– les moyennes mobiles non centrées ignorent les autres observations de la série.

© Dunod – Toute reproduction non autorisée est un délit.

6.2 Le lissage exponentiel simple (LES)

Le lissage exponentiel simple pallie les limites des moyennes mobiles non centrées. C'est une technique de prévision très simple à $t + 1$: ne dépendant que d'un seul paramètre de lissage, elle est bien adaptée pour traiter des séries temporelles qui n'ont ni tendance, ni variations saisonnières. Si le prédicteur choisi a pour horizon 1 (t+1), la méthode du lissage exponentiel simple consiste à pondérer l'observation la plus récente y_t par un coefficient α, et la prévision la plus récente $\hat{y}_{t-1}$ par le coefficient $1-\alpha$, la somme des coefficients des pondérations étant 1 :

$$\hat{y}_t = \alpha y_t + (1 - \alpha)\hat{y}_{t-1}$$

Cette formule indique que pour faire une prévision, contrairement aux moyennes mobiles, il n'est pas nécessaire de disposer d'un historique des valeurs : ce qui est très appréciable. Le coefficient α est un nombre qui varie entre 0 et 1. Si on remplace la valeur de prévision $\hat{y}_{t-1} = \alpha y_{t-1} + (1 - \alpha)\hat{y}_{t-2}$, ensuite $\hat{y}_{t-2} = \alpha y_{t-2} + (1 - \alpha)\hat{y}_{t-3}$ et ainsi de suite dans l'expression ci-dessus on obtient la formule de calcul de la prévision en fonction du coefficient α et des données historiques de la série temporelle en choisissant :

$$\hat{y}_t = \alpha y_t + \alpha (1 - \alpha) y_{t-1} + \alpha (1 - \alpha)^2 y_{t-2} + \alpha (1 - \alpha)^3 y_{t-3} + \ldots$$

Soit sous une forme plus générale :

$$\hat{y}_t = \sum_{j=0}^{t-1} \alpha \left(1-\alpha\right)^j y_{t-j}$$

On peut observer que les coefficients de pondération des données historiques α, $\alpha(1 - \alpha)$, $\alpha(1 - \alpha)^2$, $\alpha(1 - \alpha)^3$, etc. diminuent de manière exponentielle quand on remonte dans les données historiques, d'où l'appellation de lissage exponentiel. Avec cette propriété du LES d'affecter des poids plus importants aux valeurs les plus récentes, ce modèle remédie aux deux dernières limitations des moyennes mobiles. On peut démontrer que la valeur de prévision à t peut être fonction de la valeur observé en t et la prévision en $t - 1$:

$$\hat{y}_t = \alpha\, y_t + \left(1-\alpha\right)\hat{y}_{t-1}$$

En transformant cette expression on obtient une expression équivalente :

$$\hat{y}_t = \alpha\, y_t + \hat{y}_{t-1} - \alpha\, \hat{y}_{t-1} = \alpha \left(y_t - \hat{y}_{t-1}\right) + \hat{y}_{t-1}$$

L'expression $y_t - \hat{y}_{t-1}$ représente l'erreur de prévision au temps t, notée e_t. On obtient ainsi l'expression équivalente de la valeur de prévision à t :

$$y_t = \alpha\, e_t + \hat{y}_{t-1}$$

La prévision par LES, qui ne peut avoir pour horizon que t+1, est calculée avec une des formules suivantes :

$$\hat{y}_{t+1} = \alpha\, y_t + (1-\alpha)\,\hat{y}_t$$

ou

$$\hat{y}_{t+1} = \alpha\, e_t + \hat{y}_t$$

Il reste néanmoins deux problèmes à résoudre avant de pouvoir appliquer le LES. Comment calculer la toute première prévision, notée $\hat{y}_0$, puisqu'on ne dispose pas de la prévision antécédente à cette valeur ? La méthode la plus simple est de faire la moyenne arithmétique simple des deux ou trois premières valeurs observées ; mais il existe d'autres techniques plus complexes. Notons que pour une série qui comporte plusieurs données historiques, le choix de cette valeur initiale n'influence guère la prévision finale, justement en raison du coefficient de pondération faible qui sera associé à cette valeur.

Le second problème à résoudre réside dans le choix de la constante de lissage α pour que la prévision soit la plus précise possible. La méthode la plus courante est de choisir α pour qu'il minimise la somme des carrés des erreurs pour un horizon de prévision donnée. Pratiquement, il faut faire varier α entre 0,1 à 0,9 en calculant à chaque fois la somme des carrés des erreurs (*SCE*) afin de choisir la constante α qui donne la valeur minimale de ces sommes :

$$SCE = \sum_{i=1}^{t-1} \left(y_i - \hat{y}_{i-1}\right)^2$$

Utilisation du cas DOREST pour illustrer les différents aspects du LES.
Soit les données désaisonnalisées pour l'année 2015. Le tableau 6.22 donne les valeurs lissées avec un LES pour différentes constantes de prévision α. La toute première prévision est calculée comme étant la moyenne des valeurs observées en début de la série (janvier et février 2015), soit :

$$\hat{y}_0 = (8\,976 + 9\,217)/2 = 9\,096{,}5$$

Connaissant cette valeur, la prévision de janvier 2015 peut être calculée ($t = 1$):

$$\hat{y}_1 = \alpha y_1 + (1-\alpha)\,\hat{y}_0 = 0{,}1 * 8\,976 + (1\text{-}0{,}1) * 9\,096{,}5 \cong 9\,084$$

Ensuite, celle de février 2015 ($t = 2$) :

$$\hat{y}_2 = \alpha y_2 + (1-\alpha)\,\hat{y}_1 = 0{,}1 * 9\,217 + (1-0{,}1) * 9\,084 \cong 9\,098$$

Et ainsi de suite jusqu'à décembre 2015 :

$$\hat{y}_{12} = \alpha y_{12} + (1-\alpha)\,S_{11} = 0{,}1 * 9\,693 + (1-0{,}1) * 9\,259 \cong 9\,302$$

La prévision en janvier 2016 ($t + 1$) est calculée alors à l'aide de la dernière valeur observée et de la dernière prévision $\hat{y}_{12}$:

$$\hat{y}_{13} = \alpha y_{12} + (1-\alpha)\,\hat{y}_{12} = 0{,}1 * 9\,693 + (1-0{,}1) * 9\,302 \cong 9\,341$$

ou

$$\hat{y}_{t+1} = \alpha e_t + \hat{y}_t = 0{,}1 * (9\,693 - 9\,302) + 9302 \cong 9341$$

En répétant les mêmes calculs pour les autres valeurs de α, on obtient l'ensemble des autres valeurs de la série.

© Dunod – Toute reproduction non autorisée est un délit.

Tableau 6.22 – Comparaison des prévisions par LES pour différentes constantes de prévision α

Mois	Période	CA observé	Valeurs lissées avec la méthode LES				
			α = 0,1	α = 0,3	α = 0,5	α = 0,7	α = 0,9
JANVIER 2015	1	**8 976**	9 084	*9060*	9036	9012	8988
FÉVRIER 2015	2	**9 217**	9097	*9107*	9127	9156	9194
MARS 2015	3	**9114**	9099	*9109*	9121	9127	9122
AVRIL 2015	4	**9373**	9126	*9188*	9247	9299	9348
MAI 2015	5	**9 196**	9133	*9190*	9222	9227	9211
JUIN 2015	6	**9423**	9162	*9260*	9323	9364	9402
JUILLET 2015	7	**9526**	9198	*9340*	9425	9477	9514
AOÛT 2015	8	**9088**	9187	*9264*	9257	9205	9131
SEPTEMBRE 2015	9	**9470**	9215	*9326*	9364	9391	9436
OCTOBRE 2015	10	**9624**	9256	*9415*	9494	9554	9605
NOVEMBRE 2015	11	**9274**	9258	*9373*	9384	9358	9307
DÉCEMBRE 2015	12	**9 693**	9302	*9469*	9539	9593	9654
JANVIER 2016	13		***9 341***	***9536***	***9616***	***9663***	9689

La figure 6.12 qui compare la série des données prévisionnelles pour différents α et indique que le meilleur lissage est obtenu pour des valeurs faibles de cette constante. Cependant, la meilleure prédiction est faite en utilisant une constante pour laquelle la somme (moyenne) des carrés des écarts est minimale. Le calcul de ce paramètre donne, pour $\alpha = 0{,}1$:

$$SCE_{\alpha=0,1} = \sum_{i=1}^{t-1} \left(y_{i+1} - \hat{y}_i\right)^2 = \left(y_2 - \hat{y}_1\right)^2 + \left(y_3 - \hat{y}_2\right)^2 + \ldots + \left(y_{12} - \hat{y}_{11}\right)^2$$

$$SCE_{\alpha=0,1} = \left(9217 - 9084\right)^2 + \left(9114 - 9097\right)^2 + \ldots + \left(9693 - 9258\right)^2$$
$$= 763569$$

En calculant cette somme pour les autres α on trouve que la valeur minimale obtenue pour $\alpha = 0{,}3$. Sur cette base, le chiffre d'affaires prévisionnel prévu pour janvier 2016 est de 9 536 000 euros.

$$\hat{y}_{13} = 0{,}3 * 9693 + \left(1 - 0{,}3\right) * 9469 = 9536$$

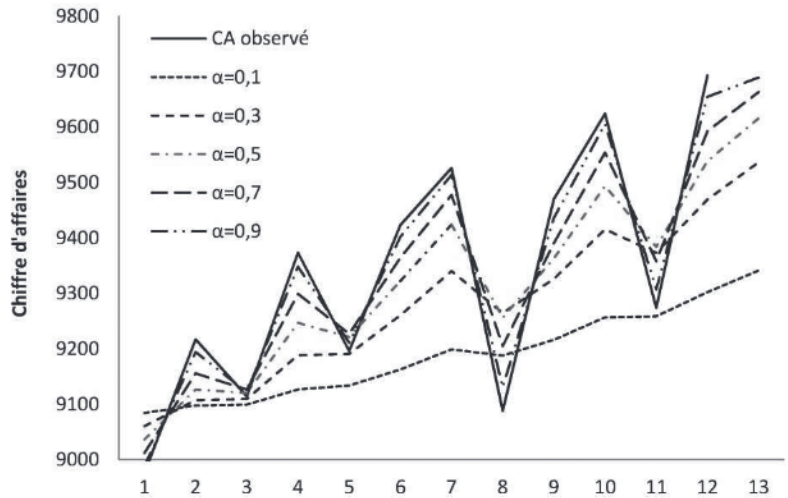


Figure 6.12 – Comparaison des prévisions par LES avec des α différents

Les techniques des moyennes mobiles et de lissages exponentiels donnent de bons résultats pour les séries temporelles plus ou moins horizontales (sans tendance) mais elles ne sont pas adaptées pour des séries ayant des tendances et des variations saisonnières marquées ou présentant des lois de variation complexes. Aussi, ces méthodes ne peuvent prédire que des valeurs à l'horizon 1. Les techniques de lissage d'ordre supérieur permettent de pallier ces problèmes, tels que la méthode des moyennes mobiles doubles, ou le lissage exponentiel double, qui ne seront abordés que brièvement dans les parties suivantes.

6.3 La prévision par les moyennes mobiles doubles d'ordre *p*

L'objectif de cette technique est de prévoir les valeurs à n'importe quel horizon, sous la forme d'une tendance linéaire de forme $a + bh$, où h est l'horizon des prévisions.

La méthode est assez simple à mettre en place et opère une série de calculs pour déterminer les deux constantes a et b :
– d'abord, le calcul des moyennes mobiles simples non centrées d'ordre p sur la série des *données brutes* ($\hat{y}'$) ;
– ensuite, le calcul des moyennes mobiles simples non centrées d'ordre p, mais cette fois sur la série des *moyennes mobiles* ($\hat{y}''$).

Pour une prévision à l'horizon h (h étant les périodes à prévoir), la méthode se résume à calculer d'abord les deux constantes a_t et b_t :

$$a_t = 2\hat{y}' - y_t'' \; ; \; b_t = \frac{1}{p-1}(\hat{y}_t' - y_t'') \text{ puis à faire l'estimation : } \hat{y}_{t+h} = a_t + b_t * \text{h.}$$

© Dunod – Toute reproduction non autorisée est un délit.

Application de cette méthode au cas DOREST sur les données désaisonnalisées en 2015.

Le tableau 6.23 résume les calculs et la figure 6.13 représente les courbes. La prévision pour janvier 2016 par les MM3 est $\hat{y}'_{t+1} = \hat{y}'_{13} = (9624+9\,274+9\,693)/3 = 9\,530$; la prévision par les MM3 doubles $\hat{y}''_{t+1} = \hat{y}''_{13}$ $(9394+9456+9\,530)/3 = 9460$. Le coefficient a se calcule comme suit : $a_t = a_{12} = 2 * 9\,530 - 9460 = 9600$; et le coefficient b ainsi : $b_t = b_{12} = 1/(3-1)(9\,530 - 9460) = 35$. Finalement, la prévision pour janvier 2016 sera : $\hat{y}_{t+1} = \hat{y}_{13} = 9600 + 35 * 1 = 9635$. Si h = 2, alors $\hat{y}_{14} = 9600 + 35 * 2 = 9670$, et ainsi de suite. La figure 6.13 trace les quatre séries sur le même espace.

Tableau 6.23 – Prévision du chiffre d'affaires par les MM doubles – Cas DOREST

Périodes	Série brute	$\hat{y}'$	$\hat{y}''$	a	b	$\hat{y}_{t+h}$
1	8 976	-	-	-	-	-
2	9 217	-	-	-	-	-
3	9 114	-	-	-	-	-
4	9 373	9 102	-	-	-	-
5	9196	9235	-	-	-	-
6	9623	9228	9188	9268	20,0	9288
7	9826	9397	9287	9507	55,0	9562
8	9088	9548	9391	9705	78,5	9784
9	9470	9512	9486	9538	13,0	9551
10	9624	9461	9507	9415	-23,0	9392
11	9274	9394	9456	9332	-31,0	9301
12	9 693	9456	9437	9475	9,5	9485
13	**-**	***9 530***	***9460***	***9600***	***35,0***	***9635***

Reste encore une difficulté : la technique des moyennes mobiles doubles d'ordre p nécessite un historique d'au moins $2p$ valeurs. La méthode du lissage exponentiel double, présentée ci-dessous, remédie à ce problème.

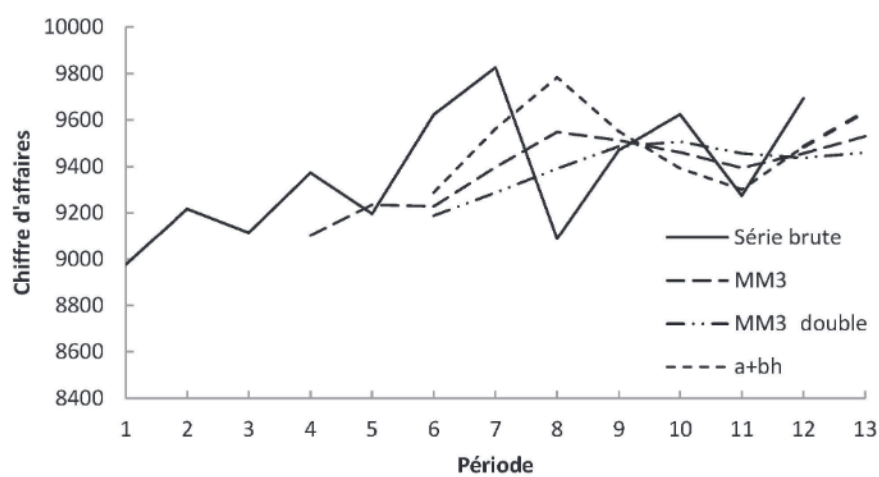


Figure 6.13 – Comparaison des prévisions par MM doubles

6.4 Le lissage exponentiel double (LED)

Il suffit de disposer de seulement trois valeurs pour utiliser le lissage exponentiel double (LED). De manière comparable à la méthode précédente, le LED consiste à appliquer un lissage exponentiel simple sur la série des données observées, puis un autre lissage exponentiel simple sur la série des données déjà lissées.

Les résultats pour le cas DOREST sont consignés dans le tableau 6.24 et la figure 6.14. La procédure du lissage exponentiel simple, telle qu'exposée au § 6.2, est appliquée sur la série des données observées. Avec une constante de prévision $\alpha = 0,3$ et $\hat{y}_0 = (8\,976 + 9\,217)/2 = 9096,5$, la première valeur de la colonne LES1 est alors obtenue à l'aide de l'expression :

$\hat{y}_1 = \alpha y_1 + (1 - \alpha)\,\hat{y}_0 = 0,3 * 8\,976 + (1 - 0,3) * 9\,096,5 = 9060$ et ainsi de suite.

Les valeurs de la colonne LES2 représentent le lissage exponentiel simple appliqué aux valeurs de la colonne LES1 selon la même procédure que précédemment. Les paramètres *a* et *b* à l'instant *t* sont calculés respectivement à l'aide des formules :

$$a_t = \frac{\alpha}{1-\alpha}(\hat{y}_t' - \hat{y}_t'') \; ; \; b_t = \hat{y}_t' - \hat{y}_t''$$

Alors la prévision d'une période par LED en *t+h* vaut : $\hat{y}_{t+h} = a_t + b_t h$.

Finalement, la prévision par LED en t+1 correspond à la somme des paramètres a et b en t. Et pour calculer les prévisions en *t+2*, *t+3*, etc., il suffit d'ajouter le paramètre *a* à la prévision précédente. Par exemple, $\hat{y}_{t+2} = \hat{y}_{t+1} + a$, $\hat{y}_{t+3} = \hat{y}_{t+2} + a$, et ainsi de suite.

Pour le cas DOREST, en décembre 2015, les coefficients *a* et *b* valent :

$$a_{12/2015} = \frac{0,3}{1-0,3}(9469 - 9370) = 42 \text{ et } b_{12/2015} = 2 * 9469 - 9370 = 9568$$

d'où on déduit la prévision de décembre 2015 : $\hat{y}_{12/2015} = 9418 + 19 = 9437$.

La prévision en janvier 2016 (t+1) est alors $\hat{y}_{1/2016} = 9568 + 42 = 9610$; et à partir de février 2016, il suffira d'ajouter 42 à la prévision précédente pour avoir la valeur correspondante.

© Dunod – Toute reproduction non autorisée est un délit.

Tableau 6.24 – Lissage exponentiel double – Cas DOREST

Mois	Période	CA observé	LES1	LES2	a	b	PREV
JANVIER 2015	1	8 976	9060	9076	-7	9044	
FÉVRIER 2015	2	9 217	9107	9085	9	9129	9037
MARS 2015	3	9114	9109	9092	7	9126	9138
AVRIL 2015	4	9373	9188	9121	29	9255	9133
MAI 2015	5	9 196	9190	9142	21	9238	9284
JUIN 2015	6	9423	9260	9177	36	9343	9 259
JUILLET 2015	7	9526	9340	9226	49	9454	9379
AOÛT 2015	8	9088	9264	9237	12	9291	9503
SEPTEMBRE 2015	9	9470	9326	9264	27	9388	9303
OCTOBRE 2015	10	9624	9415	9309	45	9521	9415
NOVEMBRE 2015	11	9274	9373	9328	19	9418	9566
DÉCEMBRE 2015	12	9 693	9469	9370	42	9568	9437
JANVIER 2016	*13*						*9610*
FÉVRIER 2016	*14*						*9653*
MARS 2016	*15*						*9695*

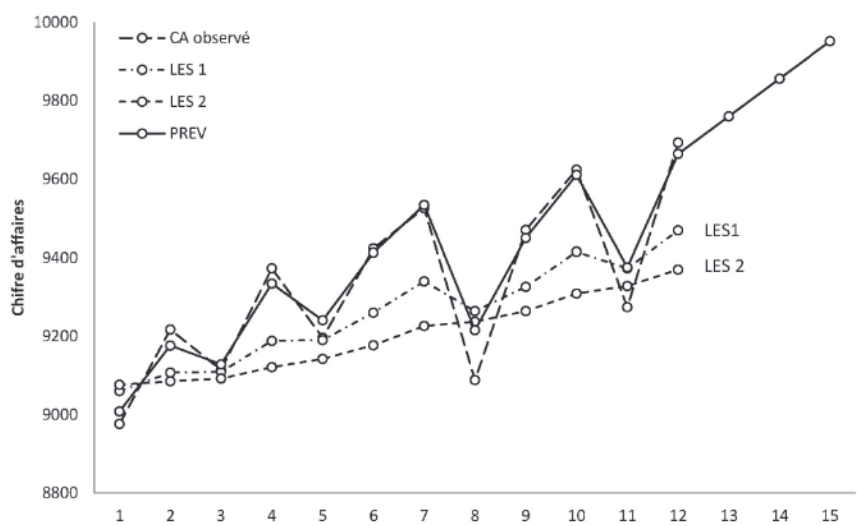


Figure 6.14 – Lissage exponentiel double – Cas DOREST

6.5 Autres méthodes de prévision par lissage

Il existe bien d'autres méthodes de prévision par lissage dont le développement sort du cadre de cette section. La Méthode de Holt, par exemple qui est adaptée pour les séries non saisonnières, représente une généralisation de la méthode de lissage exponentiel double ; la méthode de Holt-Winter, contrairement aux méthodes de lissage LES, LED et de Holt, tient compte de la saisonnalité : elle est donc adaptée pour les séries ayant des variations saisonnières.

Pour aller plus loin

Les notions développées dans ce chapitre constituent une bonne base pour comprendre l'essentiel et l'importance de l'analyse des séries temporelles. Les modèles développés dans cette section permettent de modéliser des phénomènes simples dont les caractéristiques (moyennes, variance, etc.) ne varient pas avec le temps. Dans le cas plus complexe des séries temporelles, il faut recourir à des modèles d'analyse plus élaborés. La méthodologie d'analyse des séries temporelles la plus utilisée est celle connue sous le nom de ARMA (*Auto Regressive Moving Average*) qui est un cas particulier d'un modèle plus général connu sous le nom de ARIMA (*Autoregressive Integrated Moving Average*). Dans les années 1970, les statisticiens George Box et Gwilym Jenkins ont mis en place une méthode qui intègre ces modèles pour calculer le meilleur ajustement d'une série temporelle.

L'ESSENTIEL

■ *Les techniques de prévision*

Il existe un certain nombre de modèles de prévision, plus ou moins complexes, dont certains logiciels – de traitement statistique ou plus spécialisés – permettent l'usage. L'objectif de cette section est moins d'initier à leur usage qu'à comprendre la logique intellectuelle d'une prévision qui s'appuie sur les données historiques pour prévoir à court ou moyen terme ce que peuvent être les valeurs à venir, si le futur se comporte de manière comparable au passé récent.

■ *Les composantes d'une série temporelle liées à une prévision*

Le choix d'une technique de prévision va dépendre pour partie du rôle plus ou moins déterminant de diverses composantes dans la structure des données de la série temporelle analysée : sa tendance générale (ou *trend*) qui indique le sens de son évolution – croissante, stable ou décroissante –, une éventuelle variation saisonnière (nombre d'activités économiques sont sujettes à de telles variations), voire d'autres effets, de nature cyclique, ou aléatoire. Or, pour effectuer une prévision suffisamment précise, il faut avoir identifié l'importance des composantes qui viennent perturber le *trend* afin de neutraliser leur effet, principalement celui de la variation saisonnière quand elle existe. Plusieurs techniques et différents modèles permettent de traiter ces diverses situations.

© Dunod – Toute reproduction non autorisée est un délit.

Orientations bibliographiques

A – Ouvrages généraux sur le traitement statistique des données

Biales C. et Fenneteau H. : *Analyse statistique des données, applications et cas,* Ellipses, Paris, 1993.

Bourroche J.-M. et Saporta B. : *L'analyse des données,* Presses Universitaires de France, collection « Que sais-je ? », Paris, 1980.

Carricano M, Poujol F. et Bertrandias L. : *Analyse des données avec SPSS*, Pearson, 2e édition 2010.

Coutrot B, Droesbeke F., *Méthodes de prévision*, PUF, coll. « Que sais-je ? », Paris, 1984.

Croutsche J.-J. : *Pratique de l'analyse des données en marketing et gestion,* Editions Eska, Paris, 1997.

Deroo M., Dussaix, A.-M, *Principes et analyse des enquêtes par sondages*, Presses Universitaires de France, Paris, 1980.

Evrard Y. *et alii* : *Market. Etudes et recherches en marketing. Fondements, Méthodes,* Dunod, Paris, 2000.

Giannelloni J.-L. et Vernette E. : *Etudes de marché,* Vuibert, Paris, 2001 (2e édition).

Giard V. : *Statistique appliquée à la gestion*, Economica, Paris, 1982.

Grais B. : *Méthodes statistiques,* Dunod, Paris, 1974.

Hann C. et Macé S. : *Méthodes statistiques appliquées au management,* Pearson, 2012.

Jolibert A. et Jourdan P. : *Marketing Research. Méthodes de recherche et d'études en marketing,* Dunod, Paris, 2006.

Lebaron F. : *L'enquête quantitative en sciences sociales : Recueil et analyse des données*, Dunod, coll. « Psycho sup », 2006.

Lebart L., Piron M. et Morineau A. : *Statistique Exploratoire Multidimensionnelle*, Paris, Dunod, 2006.

Liquet J.-C., Flambard S. et Jean S. : *Cas d'analyse des données en marketing,* Lavoisier, Tec et Doc, Paris, 2003.

Malhotra N., Décaudin J.-M. et Bouguerra A. : *Etudes marketing avec SPSS,* Pearson Education, Paris, 2007.

Védrine J.-P. : *Le traitement des données en marketing,* Editions d'Organisation, Paris, 1991.

Volle M. : *Analyse des données*, Economica, 1997.

B – Ouvrages spécialisés sur le traitement statistique des données

Benzecri J.-P. : *L'analyse des données,* Tome 2 : *L'analyse des correspondances,* Dunod, Paris, 1973.

Berenson M. L., Levine D. L. et Krehbiel T.C. : *Basic Business Statistics. Concepts and Applications,* Pearson Prentice Hall, Upper Saddle River (N.J), 2006 (10e édition).

Brossier G. et Dussaix A.-M. : *Enquêtes et sondages : méthodes, modèles applications, nouvelles approches,* Dunod, Paris, 1999.

Bry X. : *Analyses factorielles simples,* Economica, Paris, 1995.

Chandon J-L. et Pinson S. : *Analyse typologique,* Masson, Paris, 1981.

Cibois P. : *L'analyse factorielle,* Presses Universitaires de France, Paris, 2000.

Dodge Y. et Rousson V. : *Analyse de régression appliquée*, Dunod, 2004.

Escofier B. et Pagès J. *Analyses factorielles simples et multiples ; objectifs, méthodes et interprétation*, Dunod, Paris, 2008.

Everitt B., Landau S., Leese M : *Cluster Analysis*, Oxford, Oxford University Press (4e édition), 2001.

Gouriéroux, C. et Monfort, A. : *Séries temporelles et modèles dynamiques*. Economica, Paris,1995.

Groebner D. F. et *alii* : *Business Statistics. A decision-Making approach,* Pearson Prentice Hall, Upper Saddle River (N.J), 2001.

Hair J.F.Jr., Black W.C., Babin B.J., Anderson R.E., Tatham R.L., *Multivariate Data Analysis,* Pearson-Prentice Hall, 5e édition, 2006.

Lebart L., Morineau A. et Fénelon J.-P. : *Traitement des données statistiques,* Dunod, Paris, 1979.

Lebart, L., Morineau, A. et Warwick, K.W.: *Multivariate Descriptive Statistical Analysis, Correspondence Analysis and Related Techniques for Large Matrices*, Wiley, New-York, 1984.

Newbold, P., Carlson, W. L. & Thorne, B. : *Statistics for Business and Economics,* 7th édition, Pearson, 2010.

Saporta G., *Probabilités, analyse des données et statistiques*, Ed. Technip, Paris, 1990.

Tenenhaus M. : *Méthodes statistiques en gestion,* Dunod, Paris, 1994.

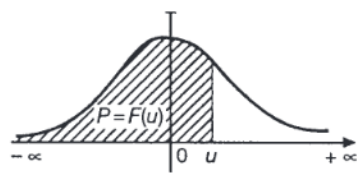
Tenenhaus M. : *Statistiques : méthodes pour décrire, expliquer, prévoir,* Dunod, Paris, 2e édition, 2007.

Tufféry S., *Data Mining et statistique décisionnelle,* Ed. Technip, Paris, 2007.

Wheelwright S.C., Makridakis S., *Choix et valeur des méthodes de prévision*, Ed. d'Organisation, Paris, 1974.

Tables statistiques

Table 1 – Fonction de répartition de la loi normale centrée réduite Probabilité $F(u)$ d'une valeur inférieure à u



u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table 2 – Fractiles d'ordre P de la loi de Student T_ν

ν \ P	0,60	0,70	0,80	0,90	0,95	0,975	0,990	0,995	0,999	0,9995
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
32	0,256	0,530	0,853	1,309	1,694	2,037	2,449	2,738	3,365	3,622
34	0,255	0,529	0,852	1,307	1,691	2,032	2,441	2,728	3,348	3,601
36	0,255	0,529	0,852	1,306	1,688	2,028	2,434	2,719	3,333	3,582
38	0,255	0,529	0,851	1,304	1,686	2,024	2,429	2,712	3,319	3,566
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,261	3,496
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
70	0,254	0,527	0,847	1,294	1,667	1,994	2,381	2,648	3,211	3,435
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
90	0,254	0,526	0,846	1,291	1,662	1,987	2,368	2,632	3,183	3,402
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Table 3 – Fractiles d'ordre P de la loi χ^2_ν

ν \ P	0,005	0,01	0,025	0,05	0,1	0,25	0,5	0,7	0,75	0,9	0,95	0,975	0,99	0,995	0,999
1	0,000	0,0002	0,001	0,004	0,016	0,102	0,455	1,074	1,32	2,706	3,841	5,02	6,635	7,88	10,827
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,408	2,77	4,605	5,991	7,38	9,210	10,60	13,815
3	0,072	0,115	0,216	0,352	0,584	1,21	2,366	3,665	4,11	6,251	7,815	9,35	11,345	13,00	16,266
4	0,207	0,297	0,484	0,711	1,064	1,92	3,357	4,878	5,39	7,779	9,488	11,14	13,277	15,00	18,467
5	0,412	0,554	0,831	1,15	1,610	2,67	4,351	6,064	6,63	9,236	11,070	12,83	15,086	16,86	20,515
6	0,676	0,872	1,24	1,64	2,204	3,45	5,348	7,231	7,84	10,645	12,592	14,45	16,812	18,65	22,457
7	0,989	1,24	1,69	2,17	2,833	4,25	6,346	8,383	9,04	12,017	14,067	16,01	18,475	20,37	24,322
8	1,34	1,64	2,18	2,73	3,490	5,07	7,344	9,524	10,22	13,362	15,507	17,53	20,090	22,03	26,125
9	1,73	2,09	2,70	3,33	4,168	5,90	8,343	10,656	11,39	14,684	16,919	19,02	21,666	23,66	27,877
10	2,16	2,56	3,25	3,94	4,865	6,73	9,342	11,781	12,55	15,987	18,307	20,48	23,209	25,25	29,588
11	2,60	3,05	3,82	4,57	5,578	7,58	10,341	12,899	13,70	17,275	19,675	21,92	24,725	26,82	31,264
12	3,07	3,57	4,40	5,23	6,304	8,43	11,340	14,011	14,85	18,549	21,026	23,34	26,217	28,35	32,909
13	3,57	4,11	5,01	5,89	7,042	9,30	12,340	15,119	15,98	19,812	22,362	24,74	27,688	29,87	34,528
14	4,07	4,66	5,63	6,57	7,790	10,16	13,339	16,222	17,12	21,064	23,685	26,12	29,141	31,37	36,123
15	4,60	5,23	6,26	7,26	8,547	11,03	14,339	17,322	18,25	22,307	24,996	27,49	30,578	32,85	37,697
16	5,14	5,81	6,91	7,96	9,312	11,91	15,338	18,418	19,37	23,542	26,296	28,85	32,000	34,31	39,252
17	5,70	6,41	7,56	8,67	10,085	12,79	16,338	19,511	20,49	24,769	27,587	30,19	33,409	35,76	40,790
18	6,26	7,01	8,23	9,39	10,865	13,67	17,338	20,601	21,60	25,989	28,869	31,53	34,805	37,19	42,312
19	6,84	7,63	8,91	10,12	11,651	14,56	18,338	21,689	22,72	27,204	30,144	32,85	36,191	38,62	43,820
20	7,43	8,26	9,59	10,85	12,443	15,45	19,337	22,775	23,83	28,412	31,410	34,17	37,566	40,03	45,315
21	8,03	8,90	10,28	11,56	13,240	16,34	20,337	23,858	24,93	29,615	32,671	35,48	38,932	41,43	46,797
22	8,64	9,54	10,98	12,34	14,041	17,24	21,337	24,939	26,04	30,813	33,924	36,78	40,289	42,83	48,268
23	9,26	10,20	11,69	13,09	14,848	18,13	22,337	26,018	27,14	32,007	35,172	38,08	41,638	44,21	49,728
24	9,89	10,86	12,40	13,85	15,659	19,03	23,337	27,096	28,24	33,196	36,415	39,36	42,980	45,59	51,179
25	10,56	11,52	13,12	14,61	16,47	19,940	24,34	28,172	29,340	34,38	37,652	40,65	44,314	46,96	52,620
26	11,20	12,20	13,84	15,38	17,29	20,840	25,34	29,246	30,430	35,56	38,885	41,92	45,642	48,32	54,052
27	11,84	12,88	14,57	16,15	18,114	21,75	26,336	30,319	31,53	36,741	40,113	43,19	46,963	49,67	55,476
28	12,49	13,56	15,31	16,93	18,939	22,66	27,336	31,391	32,62	37,916	41,337	44,46	48,278	51,02	56,893
29	13,15	14,26	16,05	17,71	19,768	23,56	28,236	32,461	33,71	39,087	42,557	45,72	49,588	52,36	58,302
30	13,82	14,95	16,79	18,49	20,599	24,48	29,336	33,530	34,80	40,256	43,773	46,98	50,892	53,70	59,703
40	20,73	22,16	24,43	26,51	29,05	33,66	39,34	44,16	45,62	51,80	55,76	59,34	63,69	66,78	73,44
50	28,01	29,71	32,36	34,76	37,69	42,94	49,33	54,72	56,33	63,17	67,50	71,42	76,15	79,50	86,69
60	35,55	37,48	40,48	43,19	46,46	52,29	59,33	65,23	66,98	74,40	79,08	83,30	88,38	91,96	99,63
70	43,29	45,44	48,76	51,74	55,33	61,70	69,33	75,69	77,58	85,53	90,53	95,02	100,42	104,22	112,34
80	51,18	53,54	57,15	60,39	64,28	71,14	79,33	86,12	88,13	96,58	101,88	106,63	112,33	116,33	124,86
90	59,21	61,75	65,65	69,13	73,29	80,62	89,33	96,52	98,64	107,56	113,14	118,14	124,12	128,31	137,22
100	67,34	70,06	74,22	77,93	82,36	90,13	99,33	106,91	109,14	118,50	124,34	129,56	135,81	140,18	149,46

© Dunod – Toute reproduction non autorisée est un délit.

Table 4 – Distribution du F de Fisher-Snedecor pour α : 0,05

v_2	v_1													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**20**	**30**	**40**	**100**
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	248,01	250,10	251,14	253,04
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,45	19,46	19,47	19,49
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,66	8,62	8,59	8,55
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,80	5,75	5,72	5,66
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,56	4,50	4,46	4,41
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,87	3,81	3,77	3,71
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,44	3,38	3,34	3,27
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,15	3,08	3,04	2,97
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	2,94	2,86	2,83	2,76
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,77	2,70	2,66	2,59
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,65	2,57	2,53	2,46
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,54	2,47	2,43	2,35
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,46	2,38	2,34	2,26
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,39	2,31	2,27	2,19
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,33	2,25	2,20	2,12
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,28	2,19	2,15	2,07
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,23	2,15	2,10	2,02
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,19	2,11	2,06	1,98
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,16	2,07	2,03	1,94
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,12	2,04	1,99	1,91
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,10	2,01	1,96	1,88
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,07	1,98	1,94	1,85
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,05	1,96	1,91	1,82
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,03	1,94	1,89	1,80
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,01	1,92	1,87	1,78
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	1,99	1,90	1,85	1,76
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	1,97	1,88	1,84	1,74
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	1,96	1,87	1,82	1,73
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	1,94	1,85	1,81	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	1,93	1,84	1,79	1,70
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11	1,88	1,79	1,74	1,63
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,84	1,74	1,69	1,59
45	4,06	3,20	2,81	2,58	2,42	2,31	2,22	2,15	2,10	2,05	1,81	1,71	1,66	1,55
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,78	1,69	1,63	1,52
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,75	1,65	1,59	1,48
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,02	1,97	1,72	1,62	1,57	1,45
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,70	1,60	1,54	1,43
90	3,95	3,10	2,71	2,47	2,32	2,20	2,11	2,04	1,99	1,94	1,69	1,59	1,53	1,41
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,68	1,57	1,52	1,39

Table 4 (suite) – Distribution du F de Fisher-Snedecor pour α : 0,001

ν_2	ν_1													
	1	2	3	4	5	6	7	8	9	10	20	30	40	100
1	4052	4999	5403	5624	5763	5858	5928	5981	6022	6055	6208	6260	6286	6334
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,45	99,47	99,47	99,49
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	26,69	26,50	26,41	26,24
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,02	13,84	13,75	13,58
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,55	9,38	9,29	9,13
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,40	7,23	7,14	6,99
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,16	5,99	5,91	5,75
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,36	5,20	5,12	4,96
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	4,81	4,65	4,57	4,41
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,41	4,25	4,17	4,01
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,10	3,94	3,86	3,71
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	3,86	3,70	3,62	3,47
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,66	3,51	3,43	3,27
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,51	3,35	3,27	3,11
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,37	3,21	3,13	2,98
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,26	3,10	3,02	2,86
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,16	3,00	2,92	2,76
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,08	2,92	2,84	2,68
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,00	2,84	2,76	2,60
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	2,94	2,78	2,69	2,54
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	2,88	2,72	2,64	2,48
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	2,83	2,67	2,58	2,42
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	2,78	2,62	2,54	2,37
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	2,74	2,58	2,49	2,33
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,70	2,54	2,45	2,29
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,66	2,50	2,42	2,25
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,63	2,47	2,38	2,22
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,60	2,44	2,35	2,19
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,57	2,41	2,33	2,16
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,55	2,39	2,30	2,13
35	7,42	5,27	4,40	3,91	3,59	3,37	3,20	3,07	2,96	2,88	2,44	2,28	2,19	2,02
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,37	2,20	2,11	1,94
45	7,23	5,11	4,25	3,77	3,45	3,23	3,07	2,94	2,83	2,74	2,31	2,14	2,05	1,88
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70	2,27	2,10	2,01	1,82
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,20	2,03	1,94	1,75
70	7,01	4,92	4,07	3,60	3,29	3,07	2,91	2,78	2,67	2,59	2,15	1,98	1,89	1,70
80	6,96	4,88	4,04	3,56	3,26	3,04	2,87	2,74	2,64	2,55	2,12	1,94	1,85	1,65
90	6,93	4,85	4,01	3,53	3,23	3,01	2,84	2,72	2,61	2,52	2,09	1,92	1,82	1,62
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,07	1,89	1,80	1,60

© Dunod – Toute reproduction non autorisée est un délit.

Index

A

Ajustement
  linéaire 245
  qualité 207
Alpha de Cronbach 139
Analyse
  de la tendance 245
  de variance 45, 72, 74, 193
Analyse factorielle
  confirmatoire 116, 141
  des correspondances multiples 116, 143
  des correspondances simples 81, 93, 96, 186
  discriminante 187, 197
  en composantes principales 123, 128
  en facteurs communs 123, 128
  exploratoires 116, 124
ANOVA 75, 186, 211, 231
Approximation 42, 52

B

Boîte à moustaches 21, 31, 51

C

Chi-deux partiel 80, 88
Classification
  hiérarchique 176
  non hiérarchique 180
Coefficient
  d'aplatissement 28, 31
  d'asymétrie 28, 31
  d'autocorrélation 250
  de contingence 90, 93
  de corrélation 30, 203, 209
  de corrélation multiple 228
  de détermination 128, 134, 207, 209
    ajusté 209
  de détermination multiple 228
  de variation 28
  saisonniers 251
Coefficient de régression 205
  non standardisés 226
  standardisés 226

Colinéarité 236
Communauté 124, 134
Composantes d'une série temporelle 241
Contribution 98, 102, 109, 128, 145
Coordonnées 98, 104
Corrélation 203
Corrélogramme 250
Critère de convergence 183
CVS 252

D

Déciles 27
Décomposition de la variance 208, 228
Degrés de liberté 56
Dendrogramme 162, 178, 180
Diagramme de dispersion 21
Diagramme en tige et feuilles 23
Distance
  de Mahalanobis 173
  de Manhattan 173
  euclidienne 173, 178
Distribution
  conditionnelle 202
  d'échantillonnage 50, 213
Données illustratives 155
Données manquantes 14
Droite de régression 201, 205

E

Écart 48
Écarts saisonniers 253
Écart-type 26, 27
Échantillonnage 7
  aléas 59, 72
  distribution 50, 213
Échantillons
  appariés 44, 50, 56
  indépendants 44, 56
Erreur d'échantillonnage 48, 205
Estimation
  des coefficients de régression 225
  ponctuelle 218
Étalons 182, 183
Étendue 17, 26

F

Facteur 117
Fonction
  de régression 197
  discriminantes 197
Fractiles 27
Fréquence 26, 27

G

Guttman 81, 91

H

Histogramme 20, 30
Homoscédasticité 73, 203
Hypothèse
  alternative 48, 49, 54
  nulle 48, 54, 55

I

Indice
  de distance 173, 176
  de similarité 174
  de Ward 178
Inertie 99, 126
Intervalle de confiance 43, 56, 216, 218

K

Kaiser 128
KMO 122

Kruskal-Wallis 186
Kurtosis 31, 51

L

Lissage 255
  exponentiel double 263
  exponentiel simple 258
Loi
  binomiale 41, 52
  de Laplace-Gauss 41
  normale 20, 28, 41, 56
    centrée réduite 41, 51

M

Médiane 22, 25, 51, 65, 70
Méthode
  d'agrégation 176, 177
  de réallocation 180, 183
  des « centres mobiles » 182
  des moindres carrés 205, 207
Minimum local 180, 183
Modalité 9
Mode 25
Modèle
  additif 251
  déterministe 201
  mixte 255
  probabiliste 202
Moyenne 25, 40
Moyennes mobiles 248
  centrées 246
  doubles 261
  non centrées 255
Multicolinéarité 121

N

Normalité 30, 51
Nuage de points 202, 207
Nuées dynamiques 180, 183

P

Pouvoir explicatif 207, 227
Principe du modèle RLM 223, 224
Projection 111, 113, 157

Q

Q-Q plot 31, 51
Quartiles 22, 27

R

Recensement 7
Régression 15
  Linéaire Généralisée (PLS) 240
  linéaire multiple 221
  linéaire simple 200
  logistique 239
Regroupement en classes 17
Relation linéaire 201
Résidus 205
  distribution 238
  indépendance 239
  normalité 238
Rotation 130, 133

S

Saturation 130, 134
Scores factoriels 140, 198
Scoring 188
Série temporelle 241
Seuil
  de confiance 53
  de risque 52, 53
Shapiro-Wilk 31, 51
Signification globale du modèle 229

© Dunod – Toute reproduction non autorisée est un délit.

Signification globale du modèle linéaire 210
Skewness 28, 31, 51
Sturge 17

T

Tableau
  carré 87
  de Burt 143, 146, 148
  de contingence 81, 86, 95, 96
  de distribution 16
  disjonctif complet 147
Tendance 241
Test
  d'adéquation 42
  d'ajustement 42, 60
  de Bartlett 74
  de Cattell 129
  de Cochran 73
  de comparaison 40, 43, 48
  de Fisher-Snedecor 45, 72, 229
  de Hartley 73
  de Jarque-Bera 31
  de Kolmogorov-Smirnov 31, 69
  de Kruskal-Wallis 45, 74
  de Levene 74
  de Mann-Whitney 65
  de Shapiro-Wilk 33
  de sphéricité de Bartlett 121
  de Student 50
  d'hypothèse 48, 69
  d'indépendance 85, 92
  du signe 68
  en z 58
  non paramétrique 51
  paramétrique 50, 51
  post hoc 73
Théorème central limite 51, 60
Tolérance 237
Trace 99
Tri à plat 15

V

Valeur
  manquante 15
  ponctuelle 40
  propre 99, 127, 128
Variable 8, 9
  active 146, 186
  de proportion 11
  d'intervalles 11
  illustrative 146
  métrique 21, 30, 51
  muette 147
  nominale 10
  non-métrique 10
  ordinale 10
  qualitative 10
  quantitative 11
Variance 26, 27
  intergroupe 192
  intragroupe 72, 192
  résiduelle 72
Varimax 131
Vecteurs propres 127, 140
VIF 237

Y

Yule 17, 92