CHAPITRE 5

Le traitement des données

Le décideur peut disposer aujourd'hui d'une manne d'informations volumineuses et variées, internes et externes à l'entreprise : les mégadonnées ou *big data*, générées tant par les objets connectés que par les individus. En comparaison, les informations issues d'une étude de marché sont réduites, car associées à la résolution d'un problème circonstanciel et défini. Selon que l'on veuille traiter l'une ou l'autre de ces sources, les procédures d'exploitation ne sont pas les mêmes, bien que certains outils puissent être communs. Consacré aux études de marché, cet ouvrage se limite à la méthodologie des seules techniques utiles pour analyser les informations qualitatives et quantitatives recueillies, en réponse aux objectifs de l'étude.

Les **méthodes d'analyse de contenu** permettent d'exploiter les données textuelles : textes ou discours issus d'analyses documentaires, d'entretiens ou de questions ouvertes (section I). Les **méthodes d'analyse statistique** permettent de traiter les données numériques : nombres représentant les modalités de réponse dans les questions (section II).

I Les méthodes d'analyse de contenu

DÉFINITION

L'expression « analyse de contenu » recouvre un ensemble de techniques qui permettent d'étudier avec rigueur le contenu manifeste ou latent d'un document pour en déterminer *objectivement* les éléments significatifs.

Selon la finalité de l'étude et le corpus à exploiter – contenu d'entretiens, de questions ouvertes, verbatim des réunions de groupe, courriers, presse, documents audiovisuels (vidéos, publicités), etc. –, l'analyse de contenu sera réalisée avec des techniques et des objectifs

différents. Le **corpus** définit l'ensemble des contenus à analyser. Par exemple, dans un entretien, c'est l'ensemble constitué des mots, appelés **occurrences**. Le nombre d'occurrences définit le volume d'un corpus, dont la diversité de la **forme** (= mots différents) mesure l'étendue de son « vocabulaire ».

1. Objectifs et méthodes de l'analyse de contenu

Trois principaux objectifs concernent les études de marché.

• **Identifier le lexique utilisé.** Il s'agit d'étudier la présence et la fréquence d'utilisation des mots, ainsi que leur répartition ou localisation dans le corpus. C'est l'objet de l'**analyse lexicale**. Pour simplifier, nous incluons dans la boîte à outils de l'analyse de contenu, prise au sens large, les techniques de la statistique textuelle (aussi appelée analyse textuelle, lexicométrie ou textométrie).

EXEMPLE

Avec une analyse lexicale (statistique textuelle), on a pu déterminer le lexique (vocabulaire) utilisé par certaines marques automobiles pour se différencier de leurs concurrentes.

• **Identifier une thématique.** Il s'agit de dégager les principaux thèmes utilisés (par un segment de clientèle, par exemple) pour parler d'un produit ou d'une marque, puis de quantifier leur présence et leur répartition (grâce aux « unités lexicales », regroupées en « champs lexicaux »). C'est l'objet de l'**analyse thématique**.

EXEMPLE

Une étude sur la publicité automobile révèle que le thème principal sur lequel communiquent les marques de voitures de sport est le plaisir de conduire, en utilisant dans leurs brochures publicitaires des mots comme « performance », « puissance », « plaisir », « expérience de conduite », « tenue de route », « adrénaline », etc.

• **Dégager la structure ou l'organisation d'un discours.** On souhaite comprendre la manière dont se construit le sens, en étudiant notamment les relations de voisinage (réseaux de co-occurrence) entre les mots ou les signes. On y parvient avec des méthodes relevant de l'**analyse du discours** ou de la **sémiotique**.

EXEMPLE

À partir d'un échantillon d'annonces médicales, on cherche à savoir si, derrière la diversité des créations, « se cachent » des structures de communication traduisant des types de représentation de la pratique médicale auxquels les médecins seraient plus ou moins sensibles.

Dans la pratique des études de marché, on utilise principalement trois techniques d'analyse de contenu ou d'analyse linguistique.

• L'**analyse lexicale**, surtout utilisée dans le traitement des questions ouvertes, étudie de manière statistique le vocabulaire présent dans un discours, en considérant les mots séparément en tant qu'unités lexicales simples (mais aussi dans des unités lexicales complexes : expressions idiomatiques, associations de deux mots ou « collocations », réseaux de co-occurrents, etc.).

• L'**analyse thématique**, telle qu'utilisée dans le traitement des entretiens par exemple, vise à réorganiser au sein de catégories (ou « champs lexicaux ») les données textuelles recueillies pour fournir un condensé exact de l'information initiale (dimension énonciative) et en trouver la signification (dimension compréhensive et interprétative). Pour déterminer ces catégories qui matérialisent les thèmes abordés dans un corpus donné (verbatim d'entretiens), on s'intéresse principalement à la fréquence des « mots » (à leur effectif, *i.e.* le nombre d'occurrences d'un mot). Ces fréquences déterminent les **spécificités lexicales**, positives ou négatives, c'est-à-dire les mots sur- ou sous-employés dans un corpus (LEBART et SALEM, 1994).

• L'**analyse sémiotique** étudie comment la manière d'organiser les signes dans un document engendre une signification déterminée. L'usage de modèles d'analyse portant sur les structures élémentaires de la signification, comme le carré sémiotique ou le schéma narratif, exige une compétence spécifique dont l'exposé dépasse le cadre de cet ouvrage.

2. Les étapes d'une analyse de contenu catégorielle

Le chargé d'étude, à la fois codeur et analyste, dispose aujourd'hui de logiciels puissants pour conduire une analyse de contenu. Néanmoins, ces outils exigent des choix dont vont dépendre la qualité de l'analyse et sa pertinence interprétative. C'est pourquoi il est nécessaire

de rappeler les fondements de la démarche. Une analyse de contenu, lexicale ou thématique, se déroule en trois phases – préparation, codification, exploitation – comprenant chacune plusieurs étapes.

Phase de préparation

1. **Constituer le corpus** et transcrire intégralement les contenus recueillis (verbatim) dans un fichier de données (intégrant silences et hésitations). Une lecture « flottante » de l'ensemble permet de repérer les principales idées et d'envisager certaines hypothèses concernant la construction du discours.

   EXEMPLE

   On réalise une étude pour identifier et analyser les représentations et les attitudes envers le café. Le corpus est constitué du contenu intégral des trente entretiens centrés qui ont été réalisés.

2. **Définir les unités d'analyse** à la base du découpage et du codage du corpus exploité. Le codage consiste à transformer les données brutes en unités qui résument les éléments pertinents du contenu analysé.

Ce découpage repose sur des mots, expressions, phrases ou thèmes, à partir desquels on peut définir :

– des unités syntaxiques centrées sur les caractéristiques grammaticales. Par exemple, différencier le verbe du substantif ;

– des unités lexicales centrées sur le vocabulaire. Par exemple, regrouper les synonymes ;

– des unités thématiques centrées sur le sens. Par exemple, distinguer « travail » et « loisirs », « douceur » et « amertume », etc.

L'étude d'un même corpus peut être effectuée en s'appuyant sur des unités d'analyse de type différent selon l'objet de l'étude ou les hypothèses formulées.

Phase de codification

3. **Élaborer la grille d'analyse** (catégorisation). Sur la base des règles précédentes, il s'agit de définir l'ensemble des catégories et sous-catégories, permettant de rendre compte de toute la richesse des contenus analysés et d'en constituer le « dictionnaire » : le mot-clé

dans une analyse lexicale, l'idée ou le thème dans une analyse thématique (exemple dans le tableau 5.1).

Dans une étude exploratoire, ces catégories sont généralement définies a posteriori sur quelques entretiens choisis aléatoirement ; les catégories sont déterminées, puis dénommées (par un mot ou une périphrase) et leur contenu défini avec précision. Mais si l'objectif est de vérifier des hypothèses, les catégories sont définies à priori en fonction du cadre théorique les ayant générées.

Tableau 5.1 – Analyse de la publicité d'un pick-up (utilitaire)

Thème principal	Thèmes secondaires (champs lexicaux)	Unités lexicales issues du corpus (verbatim)
Capacité utilitaire du pick-up : instrument de travail	1. Force, robustesse, solidité	Transport, remorquage, robuste, 4 roues motrices, couple, charge, rangement…
	2. Vocabulaire technique lié à la mécanique justifiant robustesse et performance	Moteur V6/V8 diesel, caisse, cadre, bâti, châssis, rendement écoénergétique
	3. Argumentaire non technique : rationnel, factuel, concret et pratique, adjectifs comparatifs, superlatifs	Meilleur (de sa catégorie), mieux équipé, équipement, caractéristiques, accessoires, véritable, facile, conduite souple

La validité de la codification dépend de la qualité des catégories retenues qui, selon Berelson, doivent être :

– exhaustives : toutes les unités d'analyse sont classées ;

– homogènes et exclusives : les unités d'analyse sont définies sur les mêmes critères de constitution, et chacune doit pouvoir être classée sans ambiguïté dans une seule catégorie ;

– pertinentes et objectives : leur création doit être liée au problème à étudier, et leur définition suffisamment précise pour qu'un même contenu soit classé de manière identique quel que soit le codeur.

4. **Dépouillement systématique.** Plusieurs codeurs se partagent les documents à exploiter et procèdent à la répartition des contenus entre les catégories. L'apparition de tout nouveau thème n'appartenant pas au dictionnaire constitué fait l'objet d'une concertation pour décider de sa création.

Dans une analyse thématique, le dépouillement repose sur le postulat que le codeur peut accéder au sens manifeste du contenu pour effectuer sa répartition entre les catégories. La validité de l'analyse dépend donc du travail des codeurs, à la fois de la constance de leurs affectations (fidélité intracodeur) et de la convergence de leurs points de vue dans leurs attributions (fidélité intercodeurs).

Auparavant, l'étude des questions ouvertes s'arrêtait généralement à cette phase de codage : l'analyse de leur contenu n'était qu'une phase intermédiaire dans le traitement de l'enquête pour transformer le matériel verbal en variables numériques traitées ensuite avec les mêmes outils statistiques que les questions fermées. Aujourd'hui, la puissance de calcul des logiciels de statistique textuelle facilite le traitement des réponses spontanées aux questions ouvertes : cela permet un gain de temps et d'énergie pour le codage et le traitement du corpus.

■ Phase d'exploitation

Elle vise à traiter statistiquement la totalité des informations codées.

5. **Analyser les tableaux de codage.** Cette analyse est réalisée :
   - pour chaque entretien (analyse verticale) afin d'identifier les catégories utilisées (contrôle de l'énonciation) ;
   - pour chaque catégorie sur l'ensemble des entretiens (analyse horizontale) pour étudier la fréquence de citation.

De manière habituelle, on mesure la fréquence d'apparition d'un thème (occurrence) et la fréquence d'association entre différents thèmes (co-occurrence) et parfois un indice de localisation des thèmes dans le discours si le dépouillement effectué tient compte de leur ordre d'apparition.

6. **Interpréter les résultats** et en faire la synthèse. Au-delà des résultats statistiques obtenus, l'analyste peut tenter de dégager le contenu latent (« la dimension cachée ») de la communication, susceptible d'expliquer globalement l'intégralité du discours recueilli, en s'appuyant sur des théories préexistantes (psychologiques, psychanalytiques, sociologiques…) ou en construisant son propre modèle interprétatif.

EXEMPLE

Le concept psychanalytique d'épargne psychique (toute économie d'effort intellectuel ou psychique engendre du plaisir) a été utilisé comme facteur explicatif des réactions positives et négatives émises envers un corpus d'annonces publicitaires.

Le rapport d'analyse ne peut se limiter aux seules conclusions interprétatives : à partir de l'exploitation systématique des tableaux de codage, l'analyste doit pouvoir les justifier. Ces conclusions seront d'autant plus recevables qu'elles reposent sur une démarche analytique, reproductible par le lecteur.

Les étapes de l'analyse de contenu, au sens large, sont comparables à celles réalisées avec les méthodes et logiciels de l'analyse stricte des données textuelles (LEBART et SALEM), à savoir : constitution et préparation du corpus (codage par clés, « nettoyage » et désambiguïsation) ; segmentation du corpus (découpage en unités minimales) ; analyse avec ou sans partition du corpus.

L'avènement du *text mining,* qui permet d'explorer de volumineux corpus de données textuelles, a généré de nombreux logiciels d'analyse textuelle utilisables avec un ordinateur personnel. Si les logiciels « historiques » (Alceste, Hyperbase, Lexico, NVivo) occupent toujours le marché, bien d'autres sont accessibles à titre onéreux (Coheris Analytics SPAD, XLSTAT, SPSS Text Analytics for Surveys, SAS Text Miner…), ou gratuit (Iramuteq, FactomineR, ou encore ASTADIAG, qui propose des analyses textuelles à partir d'analyses des correspondances en 3D, avec la technique de projection géodésique).

II LES MÉTHODES D'ANALYSE STATISTIQUE

Les méthodes d'analyse statistique s'appliquent sur des données numériques : les informations sont traduites en suites de nombres (les variables) correspondant aux modalités de réponses fixées.

Selon la signification donnée à ces nombres, les variables sont de nature qualitative ou quantitative. La distinction est importante pour interpréter le résultat, pour choisir les indices statistiques pertinents et utiliser les outils statistiques adaptés.

Dans une **variable qualitative** ou non métrique, les nombres ne correspondent pas à une vraie mesure : ils symbolisent un état ou une situation et se prêtent plus à des opérations logiques que mathématiques (on n'en calcule pas la moyenne). Il en existe deux types (selon l'échelle) :

– la **variable nominale** : les nombres désignent des observations qui ont des caractéristiques différentes, à l'exemple d'un matricule qui identifie un individu, un numéro de téléphone, une adresse IP, etc. ;

– la **variable ordinale** : les nombres décrivent une relation d'ordre ou de grandeur entre les modalités de réponse sans préciser la distance qui les sépare, à l'exemple de cette échelle qui qualifie un individu selon son âge, mais où les échelons sont inégaux et n'expriment que la progression d'un état : « bébé/enfant/adolescent/adulte/senior ».

Dans une **variable quantitative** ou métrique, les nombres correspondent à une mesure vraie, définissant une grandeur ; ils se prêtent à toutes les opérations mathématiques. Il en existe deux types (selon l'échelle) :

– la **variable d'intervalles** : les nombres désignent des échelons 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. Dans l'échelle de température de Celsius, le zéro marque une origine fondée sur un critère physique, mais non l'absence de chaleur (on ne peut donc affirmer que 20 °C est deux fois plus chaud que 10 °C). Les questions utilisées pour mesurer les attitudes correspondent souvent à ce type de variable.

– la **variable de proportion** complète la précédente : le zéro est un zéro absolu et correspond à l'absence du phénomène étudié (comme l'échelle de température de Kelvin où zéro signifie absence de chaleur). La question : « Combien de fois avez-vous pris l'avion au cours des trois derniers mois ? », où la réponse est numérique, est de ce type.

Une variable métrique 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 spécifiques dans ce même intervalle (le nombre de personnes composant un foyer ne peut être qu'un nombre entier).

Tout caractère étudié peut être analysé indépendamment des autres (analyse univariée), associé à un autre (analyse bivariée) ou associé simultanément à plusieurs autres (analyse multivariée). Ces modes d'analyse reposent sur une diversité d'outils statistiques dont on trouvera une présentation accessible dans Caumont & Ivanaj [2017], Giannelloni & Vernette [2019], Delacroix *et al.* [2021].

Chaque mode d'analyse vise des objectifs spécifiques : dans une perspective descriptive, fournir une information synthétique sur l'ensemble des caractères étudiés pour en tirer des conclusions généralisables ; dans une perspective explicative, trouver les raisons rendant compte de la dispersion des résultats obtenus.

1. L'analyse univariée

DÉFINITION

Base du **tri à plat**, c'est l'étude systématique de chaque variable à l'aide d'indices statistiques.

L'analyse univariée remplit plusieurs fonctions complémentaires.

1. Contrôler la qualité des données recueillies qui doivent être exemptes de toute anomalie (réponses manquantes, réponses multiples à la place d'une réponse unique, etc.) ; et supprimer les données aberrantes.
2. Décrire les variables avec des indices statistiques qui résument toute l'information recueillie et qui peuvent être présentés de manière graphique (diagrammes) ou numérique (chiffres). On distingue (tableau 5.2) :
   - les **indices de tendance centrale** (médiane, mode, moyenne) qui fournissent un résumé de l'ensemble des mesures effectuées ;
   - les **indices de dispersion** (étendue, fractiles, variance, écart type) qui renseignent sur la répartition des informations – leur « distribution » – entre les valeurs de la variable analysée. Nombre de ces indices ne concernent que les variables métriques ou ordinales ; ils permettent de savoir comment les informations se regroupent autour de la valeur centrale : une dispersion faible traduit une homogénéité des observations, et forte, une hétérogénéité qui peut justifier une segmentation. La **fréquence** renseigne sur la dispersion des observations entre les modalités

d'une variable non métrique ; la dispersion est maximale quand toutes les modalités ont la même fréquence.

3. Vérifier la structure d'un échantillon en comparant la distribution des valeurs observées à une distribution théorique attendue (**test d'ajustement**). C'est le cas dans un sondage, pour vérifier si la structure de l'échantillon obtenu est conforme à la distribution des quotas fixés.
4. Généraliser les résultats d'un échantillon à ceux de la population (**extrapolation par estimation**). On peut estimer (avec une marge d'erreur acceptable) dans quelle mesure les valeurs observées (fréquence, moyenne) sur l'échantillon correspondent à ce qu'on obtiendrait en interrogeant toute la population. Le calcul d'un **intervalle de confiance** sur la valeur obtenue dans l'échantillon donne une idée de la fourchette à l'intérieur de laquelle la valeur de la population peut varier.
5. Comparer les résultats obtenus entre des échantillons avec un **test de comparaison** pour décider si l'écart constaté sur la variable étudiée est bien significatif d'une réelle différence.

EXEMPLE

Apprécie-t-on de la même manière une nouvelle recette selon que l'on est consommateur régulier ou non du produit ? On compare deux échantillons de mesure exclusifs, dits indépendants, car la mesure est effectuée sur des groupes définis par une caractéristique qui les différencie (consommateur régulier ou non).

La consommation de protéines « prise de masse » produit-elle les effets positifs espérés ? On compare les mensurations avant et après traitement sur les mêmes personnes. L'effet réel dépend de l'écart entre les mesures avant et après consommation. Cette situation génère deux échantillons dont les mesures sont liées (avant et après), appelés échantillons appariés.

Le tableau 5.3 présente les principaux tests statistiques utilisables pour l'analyse univariée.

2. L'analyse bivariée

DÉFINITION

Le tri croisé permet d'étudier les relations entre deux variables en croisant leurs modalités. Il génère un tableau croisé appelé **tableau de contingence**.

Tableau 5.2 – Les différents types de variables et leurs statistiques descriptives

Variables			*Indices statistiques*		
Nature	*Échelle*	*Forme*	*Valeur centrale*	*Dispersion*	*Liaison*
	nominale	Catégorie ex. : *masculin/féminin*	mode	fréquence	coefficient de contingence
Non métrique	ordinale	Rang/ordre ex. : *classement ou préférence*	mode médiane	étendue fréquence fractile	coefficients de corrélation – de Spearman (ρ) – de Kendall (τ)
Métrique	intervalle et rapport	ex. : *échelle de Likert* ex. : *âge*	mode médiane moyenne	étendue/fractile variance écart type	coefficient de corrélation de Pearson

Définition des indices statistiques

Mode : valeur de la variable recueillant le plus d'observations.

Médiane : valeur de la variable divisant l'ensemble des observations en deux groupes égaux.

Moyenne : valeur résumant l'ensemble des mesures d'une variable métrique : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Fréquence : proportion des observations présentant une modalité de la variable ($f = n/N$), avec n = effectif de l'échantillon et N = effectif de la population.

Étendue : écart entre les valeurs extrêmes observées sur la variable.

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'effectif égal.

Variance : $[\sigma^2]$: $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2$: 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.

Écart type : $\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2}$: même information que la variance mais exprimée dans la métrique de la variable.

Corrélation : valeur [comprise entre + 1 et – 1] mesurant le degré de liaison existant entre deux variables ; zéro indiquant l'absence de lien ;

coefficient de corrélation de Pearson : $r = \sum_{i=1}^{k} (x_i - \overline{x})(y_i - \overline{y})/(\sigma_x \sigma_y)$; ou coefficients de rang : tau de Kendall (τ), rhô de Spearman (ρ).

Tableau 5.3 – Classification des méthodes pour l'analyse univariée

Échantillons d'observation		Nature des variables			Types de problématique statistique
Nombre	*Type*	*Non métrique*		*Métrique : intervalle et rapport*	
		nominale	*ordinale*		
un seul		Chi – deux [χ^2] test binomial test du signe	Chi – deux [χ^2] test de Kolmogorov-Smirnov	Intervalle de confiance test de la loi normale test de Student	AJUSTEMENT : comparaison à une distribution théorique ESTIMATION de la valeur d'un indice $(f, \bar{x})$ sur la population
deux	indépendants	Chi – deux [χ^2] test de Kolmogorov test de Fisher	Chi – deux [χ^2] test de la médiane test U de Mann-Whitney	test de la loi normale test de student	COMPARAISON des indices (fréquence, moyenne, variance) entre deux échantillons
	appariés	test de Mc Nemar	test du signe test de Wilcoxon	test t sur la moyenne des différences	
trois et plus	indépendants	Chi – deux généralisé [χ^2] test de Kolmogorov	test de la médiane test H de Kruskal-Wallis	analyse de la variance	COMPARAISON des indices entre plusieurs échantillons
	appariés	test Q de Cochran	test de Friedman	analyse de la variance	

Le choix des croisements ou des tris à effectuer entre paires de variables dépend des questions d'étude à traiter ou des hypothèses préalablement formulées. De ce fait, ils sont indispensables à l'étude.

• Analyser le degré d'association entre deux variables. Divers **indices de liaison** permettent de mesurer et de tester statistiquement l'intensité et le sens de la relation. Ces indices sont différents selon la nature métrique ou non des variables (voir tableau 5.2), mais sont tous calibrés pour varier dans l'intervalle {−1, 0, +1}, 0 indiquant une absence de liaison (indépendance entre les deux variables), et 1 une liaison maximale, le signe informant sur le sens de la variation (+ : même sens ; − : sens inverse).

• Tester l'indépendance entre deux variables. L'application d'un **test d'indépendance** (χ^2 = Chi deux) sur un tableau croisé permet de savoir si les distributions respectives des deux variables étudiées sont liées sans préjuger de l'éventuelle influence de l'une sur l'autre.

EXEMPLE

Il existe un lien significatif entre la lecture d'un magazine et la profession de son lecteur si la valeur de l'indice calculé (χ^2) est supérieure à une valeur seuil définie dans la table de distribution du χ^2. Dans l'affirmative, les deux variables liées sont dites « dépendantes ».

• Élaborer un modèle formalisant les relations entre deux variables. C'est le cas de la régression (linéaire pour des variables métriques : $Y = aX + b$), souvent réalisée dans une perspective prévisionnelle.

EXEMPLE

Connaissant la surface d'un appartement (X), on va essayer de prévoir son prix (Y), sachant qu'il y a un prix minimal fixe à payer lié aux frais d'acquisition (« b », appelé constante) ; « a » est le coefficient de régression (lié au prix du m²) qui permet d'ajuster le modèle prévisionnel pour les données étudiées.

L'analyse bivariée permet d'analyser les relations de concomitance qui existent entre deux variables. Néanmoins, l'analyste procède souvent à une hiérarchisation des variables en distinguant une **variable « explicative »** qui est définie comme étant la source (la « cause ») des variations d'une **variable « expliquée »**. En toute logique, c'est bien la surface de l'appartement qui influence son prix, la profession qui détermine la lecture d'un magazine, et non l'inverse. Mais l'analyse bivariée ne permet

pas de faire l'imputation causale : c'est l'analyste qui prend la liberté d'interpréter les résultats dans une perspective causale.

3. L'analyse multivariée

DÉFINITION

Étude systématique des relations entre plusieurs variables ou groupes de variables considérées simultanément, effectuée dans une perspective soit descriptive, soit explicative.

La perspective descriptive

Son objectif est d'examiner l'ensemble des interrelations entre toutes les variables étudiées pour dégager, si elle existe, une « structure organisatrice » sous-jacente permettant de rendre compte et de comprendre la logique des liens constatés. On peut distinguer trois applications principales :

– structurer les variables en des sous-ensembles homogènes identifiables, et fournir une information condensée résumant objectivement l'information initiale pouvant aller jusqu'à créer, pour chaque sous-ensemble, une variable composite qui synthétise son contenu.

EXEMPLE

On mesure l'appréciation de plusieurs formules d'un café en dosette individuelle avec une grille de vingt items. Une analyse factorielle regroupe ces items autour de trois axes au contenu homogène : goût, intensité, arôme. Ces axes remplacent les vingt items pour comparer les cafés testés.

– extraire et sélectionner les variables les plus représentatives du contenu étudié. C'est souvent la procédure suivie au terme d'une pré-étude pour élaborer un questionnaire limité à quelques items significatifs ;

EXEMPLE

Dans la perspective d'élaborer un outil pour classer les voyageurs en fonction de leurs attitudes à l'égard du voyage en train, une pré-étude a permis d'identifier près de 150 items liés à ce thème. Une double analyse, factorielle et typologique, a permis de sélectionner 28 items pour bâtir un questionnaire opérationnel.

Tableau 5.4 – Classification des méthodes pour l'analyse multivariée

Nombre et nature des variables				Types de méthodes
Les méthodes descriptives				
qualitatives nominales quantitatives discrètes				deux variables : analyse factorielle des correspondances (AFC) plusieurs variables : analyse des correspondances multiples (ACM)
qualitatives ordinales				analyse des similarités et des préférences (MDS pour *Multi Dimensional Scaling*)
quantitatives continues				analyse factorielle en composantes principales (ACP) analyse factorielle en facteurs communs et spécifiques (AFCS)
qualitatives et quantitative				typologie (classification hiérarchique, nuées dynamiques, etc.)
Les méthodes explicatives				
variable dépendante ou expliquée		variable indépendante ou explicative		
nombre	nature	nombre	nature	
une	métrique	une/plusieurs	métriques	régression simple/régression multiple/régression généralisée
plusieurs	métriques	plusieurs	métriques	analyse canonique
une	nominale	plusieurs	métriques	analyse discriminante/arbres de décision
une	binaire	plusieurs	métriques ou non	modèle Probit/modèle Logit/régression logistique
une	ordinal	plusieurs	métriques	analyse monotone de la variance (MONANOVA)
une/plusieurs	métriques	une/plusieurs	non métriques	analyse de variance univariée (ANOVA)/multivariée (MANOVA)
une/plusieurs	métriques	plusieurs	métriques et non	analyse de la covariance univariée (ANCOVA)/multivariée (MANCOVA)
une	ordinale	plusieurs	non métriques	analyse des mesures conjointes («Trade-off»)
plusieurs	tous types	plusieurs	tous types	équations structurelles (LISREL, PLS, AMOS) et réseaux neuronaux (RNA)
plusieurs	non métriques	plusieurs	non métriques	modèles Log-linéaires
une	métrique ou non	plusieurs	non métriques	modèles de segmentation (Belson, AID, CHAID, CART, théorie de l'information)

– vérifier ou tester des hypothèses. On suppose que certaines variables doivent se trouver associées dans certaines structures préétablies ; l'application de la méthode permet de confirmer ou non l'hypothèse (on parle d'analyse confirmatoire). Cette procédure est le plus souvent utilisée pour construire les échelles de mesure multidimensionnelles.

EXEMPLE

Pour mesurer l'attitude à l'égard de la publicité, on crée une échelle fondée sur deux pôles : un pôle cognitif et l'autre affectif. On contrôle que tous les items utilisés n'appartiennent bien qu'à un seul de ces deux pôles.

Les méthodes factorielles et les techniques de classification répondent à cette double logique structurante et réductrice. Ces deux classes de méthodes sont fréquemment associées dans un protocole d'étude, les premières permettant de construire les profils de similarité nécessaires pour classer ensuite de manière homogène les observations effectuées. La diversité des algorithmes (voir tableau 5.4) est liée à la nature des variables sur lesquelles les appliquer.

1. **Les méthodes factorielles.** Elles exploitent des matrices dont les valeurs expriment l'existence, la force et la direction des relations entre toutes les variables prises deux à deux, soit : des matrices de corrélation, des tableaux de contingence, des tableaux de distances.

   Elles fournissent des systèmes d'axes (aussi appelés facteurs, dimensions ou composantes) qui structurent les informations initiales, et permettent de construire des cartes perceptuelles (ou *mapping*) représentant simultanément les variables et les observations dans un même système d'axes. Ces méthodes sont utilisées en marketing pour l'étude de l'image de marque et du positionnement.

L'analyse factorielle des correspondances simple (AFC) traite des tableaux de contingence composés de deux variables non métriques ou discrètes dont les modalités sont nombreuses.

L'analyse des correspondances multiple (ACM) traite des tableaux de contingence multiple (ou « tableaux de Burt »), composés de plusieurs variables non métriques ou discrètes.

L'analyse des similarités et des préférences (MDS) traite une matrice de distances objectives ou subjectives entre des objets (produits,

marques, voire individus) à partir de leur degré de similarité perçue par les répondants, ou de leur distance à un idéal implicite (analyse des préférences). Les données recueillies sont généralement de nature ordinale, mais elles peuvent être métriques. Il existe de nombreux algorithmes dérivés du modèle de base (Torsca, Mdscal, etc).

L'analyse factorielle exploratoire en composantes principales (ACP) ou en facteurs communs et spécifiques (AFCS) permet d'analyser la matrice des corrélations établies d'un ensemble de variables métriques.

2. **Les méthodes de classification ou de typologie**. Ces méthodes visent à former des groupes d'individus ou d'objets aussi peu nombreux et aussi homogènes que possible en fonction des similarités entre les observations établies sur la base des variables initiales ou de variables construites (comme les facteurs issus d'une analyse factorielle).

   Elles doivent satisfaire à une double contrainte de condensation et d'homogénéité, en constituant des groupes de taille variable au sein desquels les éléments constitutifs sont aussi semblables que possible (faible variabilité intragroupe), et entre lesquels ils sont aussi dissemblables que possible (forte variabilité intergroupe). En marketing, ces méthodes sont particulièrement utiles pour étudier la segmentation d'un marché et réfléchir aux stratégies de ciblage.

Il existe une grande variété d'algorithmes applicables à tous types de variables métriques ou non, regroupés dans deux classes.

• **Les modèles hiérarchiques** construisent une partition de l'ensemble des observations en établissant des liens hiérarchiques entre les groupes, liens visualisés par un arbre de classification (dendrogramme). Le nombre de classes y est déterminé a posteriori, après analyse des résultats. La classification hiérarchique ascendante (CHA) est l'un de ces modèles.

• **Les modèles non hiérarchiques** fournissent des classes indépendantes construites par agrégation directe des observations dont le profil est similaire. Ils permettent de traiter facilement de gros fichiers mais ils nécessitent de fixer à priori le nombre de classes voulu. K-Means et Nuées Dynamiques sont les deux algorithmes les plus utilisés.

Méthodes hiérarchiques et non hiérarchiques sont souvent associées dans une même étude, pour définir le nombre de classes avec les premières, puis affiner la typologie avec les secondes.

La perspective explicative

Son objectif est de mettre en évidence comment la variation des valeurs d'une variable (ou de plusieurs) détermine ou « explique » une variation des valeurs d'autres variables. Cette logique explicative consiste donc à étudier des relations statistiques de dépendance entre des groupes de variables structurés à priori en deux ensembles hiérarchisés : le groupe des **variables indépendantes** (aussi appelées « variables explicatives » ou « facteurs ») dont les variations rendent compte de celles du groupe des **variables dépendantes** (« variables expliquées », « critères »). Divers modèles couvrent cette perspective (tableau 5.4) selon la nature et le statut des variables.

EXEMPLES

La notoriété d'une marque en lancement dépend de la durée de la campagne publicitaire et de la pression médiatique exercée.

Le prix d'un appartement est fonction de sa surface, de l'étage, du quartier.

« Notoriété » et « prix de l'appartement » sont les variables expliquées, dont la variation dépend de celle des autres variables, qui sont explicatives.

La perspective « explicative » recouvre en fait trois types d'objectifs que les termes « relation », « prédiction » et « explication » peuvent résumer. Mais les deux premiers sont très étroitement liés.

1. L'analyse des **relations entre groupes de variables** et modèles prédictifs. L'étude de la nature des relations entre variables est une étape nécessaire pour préparer l'élaboration d'un modèle prédictif. Plusieurs objectifs lui sont assignés :
   - vérifier comment les variables « explicatives » peuvent rendre efficacement compte de la variation des variables « expliquées ». Le **test de Fisher-Snédécor** (F) permet de contrôler la significativité de la relation linéaire établie entre les deux groupes de variables analysées ;
   - vérifier et tester l'intensité de cette relation en utilisant des indices, tels que le **coefficient de détermination** (R2) qui

permet de mesurer le pourcentage de la variation initiale dont rend compte le modèle établi ;

- mesurer l'importance de la **contribution** de chaque variable explicative dans la variation des valeurs des variables expliquées (comme le coefficient de détermination partielle en régression) ;
- construire un modèle mathématique qui combine les variables explicatives rendant le mieux compte de la variation des variables expliquées, afin de disposer d'un modèle prédictif.

La systématisation de cette approche conduit à combiner l'usage des outils statistiques afin de construire des modèles pour :

- prévoir l'activité future à partir de l'activité passée (analyse des séries chronologiques), ou estimer un potentiel d'activité ;
- anticiper l'effet de certaines décisions commerciales sur les résultats commerciaux (nombre de modèles de marché-tests simulés reposent sur cette logique) ;
- prédire le comportement d'un individu à partir de son appartenance connue à un groupe caractérisé ;
- déterminer la future position concurrentielle d'une marque ou d'un produit sur un marché donné à partir de la connaissance de ses caractéristiques.

Les **modèles régressifs**, linéaires ou non, permettent d'analyser l'influence d'un groupe de variables sur une variable unique (régression multiple) ou sur plusieurs variables simultanément (analyse canonique), si elles sont toutes métriques ; sinon, on choisit parmi les modèles de régression monotone ou logistique.

EXEMPLE

On cherche à expliquer le montant de la dépense mensuelle en déplacement avec quatre variables : la distance parcourue (km), le revenu du ménage (euros), l'ancienneté de résidence (mois) et le goût pour les déplacements (note). Le modèle final retenu est celui dont la combinaison des variables rend le mieux compte de la dispersion des dépenses. Il est exprimé sous la forme d'une équation, avec dans cet exemple, deux variables :

Dépense = + 52,66 – 0,02 Revenu + 14,66 Distance avec R2 = + 0,98 et F = 206,78. L'équation reproduit 98 % de la variation des dépenses, et la valeur élevée du F (significative à un seuil de confiance supérieur à 95 %)

rend compte de la qualité linéaire de l'estimation avec les coefficients de régression retenus.

Adapté de Moscarola, 1995, p. 291

La **segmentation** permet d'expliquer une variable qualitative (ou quantitative) à partir d'un ensemble de variables qualitatives. Connaissant les modalités qui caractérisent une observation, avec cette procédure, on peut prédire sa position sur la variable expliquée. Il existe une grande diversité de méthodes de segmentation.

EXEMPLE

La fréquence de lecture d'un titre de presse peut s'expliquer par des regroupements – des segments – de caractères descriptifs du lectorat : classe d'âge, sexe, niveau de revenu, profession exercée, etc. Un individu partageant les caractéristiques d'un segment défini a une certaine probabilité de lire le titre étudié.

L'**analyse discriminante** permet d'identifier les variables quantitatives différenciant significativement des groupes constitués (variable qualitative), afin d'utiliser ensuite le modèle régressif (« fonctions discriminantes ») qui en découle pour prévoir l'appartenance d'une observation quelconque (individu ou objet) à l'un des groupes constitués, sur la base de ses seules caractéristiques (*scoring*).

EXEMPLE

Grâce à un questionnaire sur les attitudes à l'égard de la publicité, il est possible, à partir des fonctions discriminantes calculées sur un échantillon test, de déterminer si un nouvel enquêté appartient plutôt au groupe des « indifférents », des « publiphiles », ou des « publiphobes ».

Les **modèles Logit et Probit** permettent de déterminer la probabilité conditionnelle pour qu'un événement survienne ou non, étant donné la nature de la relation qui peut exister entre une ou plusieurs variables indépendantes, et une variable dépendante binaire.

EXEMPLE

On cherche à déterminer la probabilité pour qu'un automobiliste change ou non de marque de véhicule (variable binaire), lors de son prochain achat, en fonction du nombre d'achats antérieurs de la marque, de sa satisfaction actuelle, du nombre et de la nature des pannes subies, etc.

2. L'identification et l'**analyse de relations de causalité**. Au-delà des relations de concomitance propres aux objectifs précédents, on peut vouloir démontrer la nature causale des relations constatées et affirmer que les variables indépendantes influencent réellement les variables dépendantes. Cet objectif se décompose en deux branches ayant leurs propres outils d'analyse.

La première branche, la plus commune, consiste à vérifier comment les variables dépendantes peuvent varier sous l'influence des variables indépendantes, nommées ici « facteurs », pour mettre en évidence des relations causales. Mais cela implique souvent de construire un plan d'expérience (voir chapitre 3) qui organise spécifiquement le plan d'observation pour apporter la preuve du lien causal.

L'**analyse de la variance** et l'analyse de la covariance permettent, sous certaines conditions, de vérifier l'influence réelle d'un ou de plusieurs facteurs non métriques sur une seule variable quantitative (analyse univariée), sur plusieurs variables quantitatives simultanément (analyse multivariée), ou à la fois sur des variables métriques et non métriques (analyse de la covariance). Des conditions d'ordre méthodologique permettent de démontrer la causalité : chaque modalité du facteur correspond à un « traitement » (défini et contrôlé par l'analyste au sein d'un plan d'observation) que la variable dépendante subit.

EXEMPLE

Dans un contrôle publicitaire, on souhaite déterminer si le degré de pression publicitaire subi par les prospects a une influence significative sur la qualité de l'image de marque.

Le « degré de pression publicitaire » est le facteur : il est composé de quatre modalités correspondant à différents niveaux de pression, de la plus faible à la plus forte, établis à partir de la distribution des contacts. Les points d'image (mesurés sur une échelle de Likert) sont les variables expliquées (encore appelées « critères »).

L'**analyse conjointe** permet d'expliquer la structure d'une préférence ou d'un classement à partir de la combinaison des modalités d'un ensemble de variables indépendantes dont on considère les effets conjointement. Elle permet aussi de déterminer la contribution – l'utilité – de chacune des modalités de ces facteurs dans la construction de

cette préférence. Ce type d'analyse implique l'usage des plans d'expérience. La méthode du *trade-off* est un plan incomplet présentant les modalités des variables par paires, afin de déterminer les modalités d'une offre qui soient un compromis acceptable par le consommateur.

EXEMPLE

On a fait classer par ordre de préférence douze modèles automobiles caractérisés par cinq facteurs qualitatifs : le carrossage (3/5 portes), la propulsion (avant/arrière), la motorisation (essence, diesel, gpl, électrique, hybride), la cylindrée (petite/moyenne/grosse), l'accès aux équipements de sécurité (option/série). L'analyse du classement des modèles permet d'identifier : (a) la hiérarchie des critères, (b) les compromis acceptables selon différents groupes de conducteurs.

Les **modèles Log-linéaires** permettent d'analyser des tableaux de contingence multidimensionnels composés de variables qualitatives ayant plusieurs modalités.

La seconde branche de l'objectif causal consiste à tester l'hypothèse selon laquelle il existerait un système de relations causales entre des groupes de variables plus ou moins imbriqués, dont certaines peuvent être latentes, et pour lesquelles on a supposé la nature des relations. Par exemple, on souhaite vérifier que le système de valeurs du prospect influence ses décisions et oriente son comportement. La procédure consiste à vérifier ou à confirmer la pertinence d'une structure causale théorique en la confrontant à la réalité des données observées ; d'où son appellation d'**analyse confirmatoire**.

Diverses méthodes statistiques permettent de traiter cet objectif et de concilier, en même temps, approche empirique et approche théorique. Notons les modèles d'analyse des structures fondés sur la résolution d'équations structurelles simultanées, dont LISREL (*linear structural relationship*) et PLS (*partial least squares* – méthode des moindres carrés) avec l'outil Amos, sont les modèles les plus connus ; et les modèles d'analyse factorielle confirmatoire, fondés sur l'analyse des structures de covariances. Enfin, affranchis des contraintes liées à la nature métrique des variables et à la linéarité des interrelations, les **réseaux neuronaux** ont permis de développer de puissants modèles prédictifs et ont contribué à l'essor du *data mining*.

Nombre d'entreprises disposent de bases de données gigantesques issues de diverses sources, internes et externes. Les informations stockées, très hétérogènes (nombres, textes, images, sons, fichiers log, voire géolocalisation, etc.), caractérisent leur clientèle et informent sur les pratiques de cette clientèle. Dans le cadre d'un marketing individualisé, l'idée est de révéler avec des outils le permettant ce que « cachent » d'utile toutes ces données. Le *data mining* (« forage de données ») est une réponse technique à cette ambition.

Divers prestataires proposent des logiciels de *data mining* dotés d'un fort potentiel d'analyses automatisées pour explorer ces méga-données (*big data*). Ils sont ergonomiques et conviviaux, utilisables par un chargé d'études sans nécessiter de grandes compétences statistiques (mais devant parfois programmer des requêtes…). Tous les algorithmes s'appuient sur des modèles statistiques, tel qu'exposé plus haut. Mais leur usage est calibré dans un protocole d'exploration correspondant à deux orientations complémentaires.

- **Révéler** des structures dans l'organisation des données exploitées : c'est la finalité du *data mining* proprement dit, qui va permettre de mieux comprendre le comportement du client. Il repose largement sur les méthodes statistiques classiques d'analyse de structures.
- **Prédire**, et si possible en temps réel, ce que pourrait être le comportement d'un client, à l'exemple des sites marchands proposant des offres lors de la navigation d'après ses parcours et achats antérieurs. C'est le domaine du *machine learning* qui repose sur l'usage des méthodes prédictives : classification et typologie, régression et analyse discriminante (*scoring*).

Cependant, il convient de rester conscient que ces solutions ne résolvent pas toutes les questions qu'un décideur peut se poser. Aussi, les enquêtes classiques restent pertinentes pour traiter des problématiques qui impliquent de connaître la subjectivité du client : son vécu, son ressenti, etc., devant les offres qui lui sont faites.