UNIVERSITE DE LISALA

CENTRE INTERUNIVERSITAIRE DE RECHERCHE PLURIDISCIPLINAIRE (CIREP)
STATUT : UNIVERSITE PUBLIQUE
Web : www.cirep.ac.cd
Email : info@cirep.ac.cd

NOTES DE COURS DE METHODES QUANTITATIVES DE LA PLANIFICATION DE L'EDUCATION

OBJECTIFS DU COURS

Objectif général :

Le cours vise à fournir aux étudiants les compétences nécessaires pour utiliser des méthodes quantitatives dans le domaine de la planification de l'éducation, en mettant l'accent sur l'analyse des données et la prise de décisions basées sur des preuves.

Objectifs spécifiques du cours :

- Comprendre les principes fondamentaux des méthodes quantitatives utilisées dans la planification de l'éducation.
- Acquérir des compétences en collecte, analyse et interprétation des données quantitatives pertinentes pour la planification de l'éducation.
- Apprendre à utiliser des logiciels statistiques pour effectuer des analyses quantitatives dans le contexte de la planification de l'éducation.
- Explorer les différentes techniques de modélisation quantitative utilisées pour prévoir les tendances et évaluer les politiques éducatives.
- Comprendre comment utiliser les résultats d'analyses quantitatives pour formuler des recommandations et des stratégies efficaces en matière de planification de l'éducation.
- Développer des compétences en communication des résultats d'analyses quantitatives de manière claire et pertinente pour les décideurs en éducation.
- Examiner les limites et les biais potentiels des méthodes quantitatives dans le contexte de la planification de l'éducation.
- Réfléchir sur l'importance de l'éthique et de l'intégrité dans l'utilisation des méthodes quantitatives en planification de l'éducation.

Chapitre 1. NOTIONS DE BASE

Aborder avec aisance les méthodes quantitatives nécessite de comprendre et maitriser d'entrée quelquestermes et outils essentiels, d'usage récurrent. Ce premier chapitre les présente.

1.1. INDIVIDU, POPULATION, VARIABLE

L'analyse quantitative porte toujours sur un ensemble d'objets, par exemple les résultats scolaires d'un élève, ou les effectifs des classes d'un établissement, ou les salaires des enseignants, etc. Par exemple, si l'on étudieles notes d'un unique élève, l'objet n'est pas l'élève mais la note. Chacun des objets du groupe d'objets étudiés est un « individu » (ou « unité statistique »). On appelle « population » le groupe d'objets étudiés.

On appelle « variable » une relation qui, à chaque individu d'une population donnée, associe une « valeur ». Par exemple, la variable « Notes obtenues par une promotion d'étudiants à l'examen de sociologie », associe à chaque étudiant une valeur qui est la note que cet étudiant a obtenue à l'épreuve de sociologie. Ou encore, autre exemple, la variable « Diplôme universitaire initial des enseignants de Sciences de l'éducation et de la formation » associe à chaque enseignant une valeur qui est l'intitulé du premier diplôme universitaire que cet enseignant a obtenu.

1.2. INDICE

Un indice est un nombre qui permet de comparer une valeur à une référence dans le temps ou dans l'espace. Soit x une valeur à un moment donné ou à un endroit donné, soit x^* la référence à laquelle on souhaite comparer x, l'indice de x par rapport à sa référence est défini par :

$$— \quad Indice = \frac{x}{x^*}$$

Par exemple, si le nombre d'élèves inscrits est de 300 dans l'établissement de référence A, de 200 dans l'établissement B et de 400 dans l'établissement C :

- l'indice du nombre d'inscrits par rapport à la référence A est de 0,66 dans l'établissement B (200/300) ; et
- l'indice du nombre d'inscrits par rapport à la référence A est de 1,33 dans l'établissement C (400/300).

De même, si le nombre d'inscrits dans un établissementD est de 150 à une date de référence (par

exemple octobre 2016), de 250 en octobre 2017 et de 300 en octobre 2018, l'indice du nombre d'inscrits par rapportà la date de référence est de 1,66 en octobre 2017 et de2 en octobre 2018.

L'usage des indices permet une comparaison directe etsimple de valeurs différentes au moyen d'une référencecommune. Il permet aisément aussi de mesurer desécarts dans l'espace. Par exemple, l'indice 0,66 del'établissement B signifie aussi que l'effectif desinscrits de l'établissement B :

- *représente* 66% de celui des inscrits de l'établissement de A ;
- *est de 34% inférieur* à celui de l'établissement A (0,66 – 1 = 0,34) ;
- *est égal à 0,66 fois* celui de l'établissement A ;
- *a un coefficient multiplicateur de* 0,66 par rapport àl'effectif des inscrits de l'établissement A.

L'usage des indices permet tout aussi aisément de mesurer des variations au cours du temps. Par exemple,les indices 1,66 et 2 de l'établissement D signifient aussi que :

- le nombre d'inscrits a *augmenté de*
  - 66% en octobre 2017 par rapport à octobre 2016(1,66 – 1 = 0,66 = 66%) ;
  - 100% en octobre 2018 par rapport à octobre 2016(2 – 1 = 1 = 100%) ;
- le nombre d'inscrits a *été multiplié par*
  - 1,66 entre octobre 2016 et octobre 2017 ;
  - 2 entre octobre 2016 et octobre 2018.

Dans la comparaison, la valeur de référence est égale à1, donc on dit que l'indice est en base 1. Mais une autrepratique courante consiste à exprimer l'indice plutôt enbase 100 :

$$Indice\ base\ 100 = Indice\ base\ 1 \times 100$$

Ainsi par exemple, les indices 0,66 et 1,33 ci-dessus des établissements B et C deviennent respectivement 66 et
133 si on les exprime en base 100. Cela étant, l'expression en base 100 ne change pas

le principe : lesindices s'interprètent toujours par comparaison avec labase.

1.3. INDICATEURS DE STATISTIQUE DESCRIPTIVE

La statistique descriptive représente le traitement de base applicable à des données quantitatives. Elle permet de mettre en évidence les premières caractéristiques des données observées. Dans le cadre d'un compte-rendu de recherche (rapport de recherche, thèse, article), ces caractéristiques doivent faire partie de la présentation générale des données analysées.

Supposons un ensemble de données quantitatives. La statistique descriptive vise à en fournir une présentation synthétique. Cette présentation s'effectue au moyen d'*indicateurs de statistique descriptive*, qui résument les principales propriétés de ces données. Deux types d'indicateurs de statistique descriptive sont d'usage particulièrement fréquent : les *indicateurs de tendance centrale* et les *indicateurs de dispersion*.

1.3.1. Indicateurs de tendance centrale

Soit une population comprenant n individus dont on étudie la variable X . Par exemple un groupe de n étudiants dont on étudie la variable « notes à l'épreuve d'anglais ». La valeur de la variable X pour un étudiant i est notée x_i

Tableau
Notes à l'épreuve d'anglais

	Étudiants (i)	Notes à l'épreuve d'anglais (x_i)
Étudiant-e 1	Chloé	10
Étudiant-e 2	Lucas	10
Étudiant-e 3	Leo	11
Étudiant-e 4	Nathan	12
Étudiant-e 5	Paul	12
Étudiant-e 6	Noémie	13

Étudiant-e 7	Emma	14
Étudiant-e 8	Léa	14
Étudiant-e 9	Sarah	16
Étudiant-e 10	Abdel	17

Les indicateurs de tendance centrale permettent de mettre en évidence des caractéristiques représentativesde la plupart des valeurs de la série.

1.3.1.1. Fréquence

La fréquence d'une modalité traduit le nombre de fois où cette modalité est représentée dans une population. Elle est définie par le rapport :

$$\text{fréquence} = \frac{\text{nombre d'occurences de la modalité}}{\text{effectif de la population}}$$

Elle s'exprime en pourcentage. Par exemple, si 3 élèvessur les 25 que compte une classe ont obtenu la note de

17/20, la fréquence de la modalité 17/20 est $\frac{3}{25} = 0{,}12$.

Dans le Tableau 1.2, les modalités 10, 12 et 14 ont chacune une fréquence de 0,20. La fréquence est de 0,10 pour chacune des modalités 11, 13, 16 et 17. La somme des fréquences des modalités d'une série est toujours égale à 1.

1.3.1.2. Mode

Le mode d'une série est la valeur de caractère la plus fréquente dans la série. Par exemple, dans le Tableau 1.2, il y a trois modes, 10-12-14. On dit que la série (c'est-à-dire ici la série des notes à l'épreuve d'anglais)

est *trimodale*. Une série peut avoir un seul mode (*série monomodale*), deux modes (*bimodale*), plusieurs modes (*multimodale* ou *plurimodale*).

1.3.1.3. Médiane

La médiane d'une série est la valeur de caractère qui sépare la série en deux sous-groupes de même effectif. Les valeurs de la série doivent avoir été au préalable rangées par ordre, croissant ou décroissant. C'est le cas de la série du Tableau 1.2, où les valeurs sont rangées en ordre croissant. Par exemple, dans le Tableau 1.2, la médiane est 12,5. Si l'effectif n'avait comporté que les 9 premiers étudiants, la médiane aurait été 12. De façon générale :

- lorsque l'effectif de la série est impair, la médiane est la $(\frac{n+1}{2})$-ième valeur ;
- lorsque l'effectif de la série est pair, la médiane est égale à

$$\frac{(\frac{n}{2}) \text{ ième } valeur + (\frac{n}{2} + 1) \text{ ième } valeur}{2}$$

1.3.2. Indicateurs de dispersion

Les indicateurs de dispersion cherchent plutôt à souligner les éléments de variabilité et d'hétérogénéité au sein de la série.

1.3.2.1. Minimum, maximum, extremum, étendue

La valeur la plus basse (minimum) et la plus haute (maximum) d'une série en sont les extrema (par exemple 10 et 17 dans le Tableau 1.2).

L'étendue d'une série est l'écart entre son maximum et son minimum (7 dans l'exemple).

1.3.2.2. Quartiles, déciles, centiles

Lorsqu'une série a un effectif suffisamment élevé, on peut en étudier l'hétérogénéité en la décomposant en sous-groupes et en comparant les caractéristiques et comportements de ces sous-groupes.

Une méthode de décomposition usuelle consiste à découper la série en quarts, ou en dixièmes, ou en centièmes. On appelle quartiles, déciles et centiles les valeurs de caractère qui délimitent les sous-groupes.

Dans une décomposition par quarts, la série, ordonnée par valeurs croissantes (comme par exemple dans le Tableau 1.2), est découpée en quatre quarts. On définit comme suit 3 quartiles, Q_1, Q_2 et Q_3:

- Q_1, le premier quartile, est la valeur telle qu'un quart de l'effectif a une valeur de caractère inférieure à Q_1 ;
- Q_2, le deuxième quartile, est la valeur telle que la moitié de l'effectif a une valeur inférieure à Q_2. Q_2 est égal à la médiane ;
- Q_3, le troisième quartile, est la valeur telle que trois quarts de l'effectif ont une valeur inférieureà Q_3.

Avec les extrema, les quartiles constituent les bornes de chaque quart. Le premier quart est donc délimité par le minimum (borne inférieure) et le premier quartile (borne supérieure) ; le deuxième quart, par le premier quartile et le deuxième quartile ; le troisième quart, parle deuxième quartile et le troisième quartile ; et le quatrième quart, par le troisième quartile et lemaximum.

EXEMPLE 1.1

Considérons le pourcentage d'étudiants salariés dans différents pays.

Part des étudiants salariés dans la tranche des 15-29 ansdans les pays de l'OCDE (%) – Année 2011

Source : OCDE (2013), *Regards sur l'éducation*
(Annexe Tableau C5.2a).

NB : la moyenne de l'OCDE s'établit à 11%.

N°	Pays	%
1	Slovaquie	2,1
2	Grèce	2,2
3	Hongrie	2,2
4	Italie	2,5
5	Belgique	3,5
6	République Tchèque	3,6
7	Espagne	4,7
8	Corée du sud	5,2
9	Portugal	5,3
10	Turquie	5,5
11	France	5,9
12	Luxembourg	5,9
13	Chili	6,8
14	Mexique	6,9
15	Irlande	7,1
16	Pologne	7,8
17	Allemagne	8,5
18	Autriche	9,6
19	Israël	10,5
20	Estonie	10,9

N°	Pays	%
21	Suède	11,1
22	Royaume Uni	11,3
23	Suisse	11,9
24	États-Unis	15,1
25	Norvège	15,3
26	Finlande	16,0
27	Slovénie	16,9
28	Nouvelle Zélande	17,4
29	Canada	17,6
30	Australie	21,1
31	Islande	26,6
32	Danemark	32,1
33	Pays-Bas	32,4

Une méthode simple pour calculer les quartiles est celledes *médianes partielles*.

Calcul de Q_1

On considère la première moitié de la série (ou lapremière moitié à laquelle on ajoute la médiane, si l'effectif de la série est impair), et on en calcule la médiane. Ici par exemple, l'effectif de la série estimpair ($n = 33$), la médiane de l'ensemble de la série est la $17^{\text{ème}}$ valeur, soit 8,5% (pourcentage pour l'Allemagne). La première moitié de la série comprend donc les valeurs de 2,1 à 8,5. La médiane de cette première moitié (première « médiane partielle ») est la $9^{\text{ème}}$ valeur, c'est-à-dire 5,3% (pourcentage du Portugal). Cette première médiane partielle est le premier quartile, Q_1.

Calcul de Q_3

On considère à présent la seconde moitié de la série (augmentée de la médiane si

l'effectif de la série est impair), et on en calcule la médiane partielle. Dans l'exemple, la seconde moitié de la série comprend les valeurs de 8,5 à 32,4. La médiane partielle de cette seconde moitié (seconde « médiane partielle ») est la $25^{ème}$ valeur de la série, c'est-à-dire 15,3 (pourcentage de la Norvège). Cette seconde médiane partielle est le troisième quartile, Q_3.

Résultats

$$\begin{cases} Q_1 = 5{,}3 \\ Q_2 \text{ est la médiane} : Q_2 = 8{,}5 \\ Q_3 = 15{,}3 \end{cases}$$

L'identification des quarts permet ainsi de souligner les éléments d'hétérogénéité. Par exemple, alors que la moyenne de l'OCDE s'établit à 11%, on voit que les pourcentages sont de plus de moitié inférieurs à la moyenne dans le premier quart, et peuvent lui être jusqu'à deux fois supérieurs dans le quatrième.

On peut observer toutefois que les valeurs extrêmes sont parfois aberrantes (c'est-à-dire considérablement plus faibles ou plus élevées que le reste de la série) et dans ce cas influencent très sensiblement le comportement des premiers et derniers quarts. Dans l'exemple, les deux dernières valeurs de la série (Danemark et Pays-Bas) pèsent fortement sur la caractérisation du dernier quart. L'influence excessive des valeurs extrêmes conduit souvent à les écarter de l'analyse. Pour ce faire, soit on élimine de la série les valeurs extrêmes elles-mêmes, soit on élimine de la segmentation les sous-groupes qui contiennent les valeurs extrêmes. Lorsque le découpage est par quarts, éliminer les premiers et derniers quarts revient cependant à amputer la série de la moitié de son effectif, ce qui est beaucoup. On recourt alors plutôt – si l'effectif de la série est suffisamment élevé – à un découpage en dixièmes ou centièmes. On définit 9 déciles (D_1 à D_9), qui partagent la série en 10 dixièmes ; ou 99 centiles (C_1 à C_{99}), qui la partagent en 100 centièmes. On peut alors exclure les premier et dixième sous-groupes et ne conserver pour l'analyse que les 8 dixièmes restants (soit 80% de la série) ; ou exclure les premier et centième sous-groupes et ne conserver que les 98 centièmes restants (98% de la série). L'intérêt de l'opération est de disposer de

fourchettes qui donnent une idée de la dispersion de la majorité des valeurs de la série :

- 50% des valeurs de la série sont comprises entre Q_1 et Q_3 ;
- 80% des valeurs de la série sont comprises entre D_1 et D_9 ;
- 98% des valeurs de la série sont comprises entre C_1 et C_{99}.

1.3.2.3. Écart à la moyenne : écart moyen, écart-typeet variance

L'idée ici est de représenter la dispersion des valeurs de la série par une mesure de l'écart des valeurs à la moyenne de la série.

Une première méthode consiste à calculer *l'écart moyen*, qui est la moyenne des écarts à la moyenne :

$$écart\ moyen = \sum^{N} (x_i - \bar{x})$$

Soit la série décrite dans la première colonne (1). La deuxième colonne décrit les étapes du calcul de l'écart moyen.

Calcul d'écart moyen

x	$x - \bar{x}$
(1)	(2)
12	5,4
8	1,4
7	0,4
4	-2,6

2	-4,6
Moyenne = 6,6	Total = 0
	Écart moyen = 0/5 = 0

L'inconvénient de l'écart moyen est que l'écart d'une valeur à la moyenne peut être positif ou négatif. Donc les écarts des différentes valeurs se compensent aumoins en partie[7], de sorte que, *in fine*, l'écart moyen reflète mal les distances entre les valeurs et la moyenne. Dans l'exemple, l'écart moyen est nul, suggérant que toutes les valeurs sont très proches de la moyenne, alorsque ce n'est pas du tout le cas.

Une méthode plus satisfaisante consiste alors à utiliser *l'écart-type*, défini comme la moyenne des carrés des écarts à la moyenne. L'écart-type est noté σ et calculé par :

$$\sigma = \sqrt{\frac{\sum^{N} (x_i - \bar{x})^2}{n}}$$

EXEMPLE

Reprenons dans le Tableau ci-dessous la série décrite dans la colonne (1). La troisième colonne décrit les étapes du calcul de l'écart-type.

Calcul d'écart moyen et d'écart-type

X	$x - \bar{x}$	$(x - \bar{x})^2$
(1)	(2)	(3)
12	5,4	29,16
8	1,4	1,96
7	0,4	0,16
4	-2,6	6,76
2	-4,6	21,16
Moyenne : 6,66	Total = 0	Total = 59,2
	Écart moyen = 0/5 = 0	Variance = 59,2/5 = 11,84
		Écart-type $= \sqrt{11,84} \cong \pm 3,44$

Par rapport à l'écart moyen, l'avantage de l'écart-type est que, pour chaque valeur, on prend le carré de son écart à la moyenne ; il n'y a donc plus de terme négatif ni de compensation. L'écart-type reflète donc mieux les écarts à la moyenne que ne le fait l'écart-moyen, ce qu'illustre bien l'exemple.

La variance est donc la somme des carrés des écarts divisés par l'effectif. C'est le carré de l'écart-type :
$variance = \sigma^2$.

Écart-type et variance sont des indicateurs très utilisés en statistique. Du point de vue de la dispersion, chacun des deux permet de représenter de façon synthétique par un chiffre unique la tendance des valeurs de la série à s'écarter de la moyenne. Soit une série : plus sa dispersion est forte, plus son écart-type et sa variance sont élevés.

1.3.2.4. Coefficient de variation

Le coefficient de variation est parfois appelé aussi *coefficient de dispersion*. La question à laquelle le coefficient de variation permet de répondre est la suivante : comment comparer la dispersion de deux séries (ou plus) dont les moyennes sont différentes ? Par exemple, le niveau des élèves est-il homogène dans les classes de différents enseignants d'une même matière ? Ou bien : y va-t-il homogénéité dans l'originesocio-économique des élèves dans différentsétablissements ou régions ? Etc.
Le coefficient de variation est défini par le rapport écart-type / moyenne. Il permet de comparer la dispersion de séries de moyennes et écarts-types différents : la série est d'autant plus dispersée que le coefficient de variation est élevé.

1.4. DISTRIBUTION STATISTIQUE ET COURBENORMALE

On appelle « distribution statistique » (ou « distribution des fréquences ») la liste des fréquences des modalitésd'une variable.
Tableau de la distribution statistique de la variable
« Catégorie socio-professionnelle du chef de famille »

Catégorie socio-professionnelle	Effectif de la modalité	Fréquence de la modalité
1. Agriculteurs exploitants	15	0,01
2. Artisans, commerçants et chefs d'entreprise	230	0,08
3. Cadres et professions intellectuelles supérieures	402	0,13
4. Professions Intermédiaires	755	0,25
5. Employés	1090	0,36
6. Ouvriers	300	0,10

7. Retraités	121	0,04
8. Autres personnes sans activité professionnelle	87	0,03
Total	*3000*	*1*

Première partie : collecter et préparer lesdonnées

La démarche quantitative commence par la collecte des données qui retracent le phénomène à analyser. Les données susceptibles d'être analysées par approche quantitative sont de formes et natures variées.

Du point de vue de la forme, on peut distinguer entre données textuelles et données numériques.

Dans le premier cas, il s'agit par exemple de textes littéraires, de documents politiques, d'articles de presse ou de transcriptions d'entretiens. Bien qu'habituellement analysés par approche qualitative suivant des méthodes d'analyse de contenu, les textes sont aussi susceptibles, le cas échéant, d'être approchés sous l'angle quantitatif. C'est l'objet de la lexicométrie.

Les données numériques peuvent être soit quantitatives soit qualitatives (ordinales ou nominales). Elles représentent la majorité des données habituellement soumises à analyse quantitative.

Du point de vue de leur nature, les données susceptibles d'être traitées par analyse quantitative peuvent êtreréparties entre données simulées et données observées, d'une part ; et entre données perceptuelles et données factuelles d'autre part. Les données simulées sont des données fictives, créées par générateurs de données suivant des procédures plus ou moins sophistiquées. Elles servent par exemple dans les phases de mise au point ou d'essai de modèles, ou encore en appui à des illustrations pédagogiques. Au contraire, les données observées sont tirées de situations réelles et décrivent le monde réel. Les données perceptuelles expriment des opinions, par opposition aux données factuelles (par exemple démographiques ou physiques), qui décriventdes faits indépendamment des opinions.

Chapitre 2. Collecter les données

Toute donnée n'est pas bonne à prendre. L'identification des bonnes données à collecter s'effectue sur la base de la question de recherche, et en fonction du cadre conceptuel et théorique adopté pour ladite recherche. Les données à collecter doivent évidemment porter de façon précise sur l'objet de la recherche, et en respecter les conceptualisations. Elles doivent constituer la meilleure représentation empirique possible des concepts à l'œuvre dans cette recherche.

Un grand nombre de données statistiques relatives à l'éducation et à la formation existent aujourd'hui en France et à travers le monde. Elles sont produites par une multitude de sources, institutionnelles ou individuelles (recherches académiques notamment), publiques ou privées, internationales, nationales ou locales. Au niveau international, l'Organisation des Nations Unies pour l'éducation, la science et la culture (UNESCO), la Banque mondiale, l'Organisation pour la Coopération et le Développement Économiques (OCDE), ou encore l'Office statistique de l'Union européenne (Eurostat), par exemple, mettent à disposition du grand public une abondance de données statistiques sur l'éducation et la formation. Ces données sont généralement agrégées (statistiques pour l'ensemble d'un pays dans tel ou tel domaine), mais comprennent souvent aussi des *micro-données*, s'est-il- dire des données obtenues de chacun des répondants individuels lors d'enquêtes. En France, deux importants producteurs de données sur l'éducation sont le ministère de l'Éducation nationale[13] et le ministère de l'Enseignement supérieur. Au niveau local, de très nombreuses sources de données sur l'éducation et la formation existent également. Par exemple, chaque établissement d'enseignement dispose de données, et en particulier de données administratives sur ses étudiants et de données sur leurs performances académiques. Pour autant, il n'est pas rare de manquer de données appropriées lorsqu'il s'agit de faire de la recherche. Bien souvent en effet, les données disponibles sont trop générales et non adaptées à l'objet spécifique de telle ou telle recherche. En particulier, les définitions adoptées par les producteurs de données pour construire leurs indicateurs et collecter les données correspondantes ne sont pas nécessairement compatibles avec le cadre théorique et conceptuel de tel ou tel travail

de recherche.

Il n'est donc pas rare que le chercheur en éducation et formation soit contraint de collecter lui-même les données pertinentes pour sa recherche. C'est l'objet de l'enquête de terrain. Dans l'absolu, l'enquête peut prendre la forme de l'observation en immersion, du questionnaire ou de l'entretien. Seules ces deux dernières formes (et davantage le questionnaire que l'entretien) sont susceptibles de générer des données traitables par approche quantitative.

2.1. L'ENQUÊTE PAR QUESTIONNAIRE

D'un point de vue quantitatif, l'enquête par questionnaire présente l'avantage de permettre de collecter, auprès d'un grand nombre de répondants, des données textuelles et/ou numériques susceptibles d'êtretraitées par lexicométrie et/ou méthodes statistiques. Collecter des réponses par questionnaire auto- administré accessible sur internet à des milliers de répondants est aujourd'hui pratique courante. Or il parait évident que plus le nombre de répondants est élevé, plus il y a de chances que les conclusions de l'analyse soient généralisables.

De nombreux ouvrages présentent les méthodes d'élaboration de questionnaire et de conduite d'enquête, et il existe également des logiciels qui en facilitent la mise en œuvre. Sur certains thèmes, il existe des questionnaires psychométriques déjà validésqui peuvent servir de point de départ. On n'y reviendra donc pas en détail dans le présent ouvrage, consacré plutôt à l'analyse des données déjà collectées. Quelques rappels essentiels s'imposent cependant.

2.1.1. Préparation du questionnaire

La préparation d'un questionnaire s'articule en deux principales phases : la phase de conception et la phase pilote.

2.1.1.1. Conception du questionnaire

Tout d'abord, construire un questionnaire nécessite de respecter un minimum de règles. L'anonymat doit être garanti, et les règles relatives au traitement des données personnelles (y compris l'âge, la profession et toute information susceptible de permettre l'identification durépondant par croisement des données) respectées.

En principe, un questionnaire commence par une introduction qui, d'une part, présente le cadre général dans lequel le questionnaire s'inscrit, de façon que le répondant comprenne la démarche à laquelle on lui demande de participer ; et d'autre part valorise et encourage la participation du potentiel répondant à la démarche proposée.

La première partie de l'introduction (présentation) indique (brièvement) la nature, l'objet, le cadre institutionnel[18] et scientifique, les objectifs et les enjeux de la recherche, en soulignant plus particulièrement les aspects de nature à intéresser et motiver le type de répondants visé.

2.1.1.2. Phase pilote

Une fois le questionnaire conçu, la phase pilote vise à en vérifier la validité et l'intelligibilité, de façon à en améliorer la qualité et l'efficacité avant administration aux répondants. La vérification de la validité porte sur trois principaux volets : couverture du champ, fiabilité,et cohérence interne. La vérification de l'intelligibilité vise à s'assurer que le questionnaire sera lisible et compréhensible pour le public auquel il s'adresse.

2.1.1.2.1. Couverture du champ

Le questionnaire doit, tout d'abord, couvrir effectivement le champ qui lui est assigné, et n'en manquer aucun aspect essentiel. Il importe donc de vérifier que les domaines (sections thématiques) et items (questions) du questionnaire couvrent bien tous les aspects et dimensions de nature à permettre au chercheur de collecter l'information qu'il vise au travers du questionnaire. Pour traiter ce premier volet de la validité, le principe est de soumettre le questionnaire à des experts (chercheurs et

praticiens) familiers du champ, afin de recueillir leur avis sur le point de savoir si toutes les principales thématiques et dimensions sont couvertes. Il est de bonne méthode de disposer d'un critère objectif sur la base duquel on peut décider que les experts ont ou non entériné le questionnaire qui leur a été soumis, et un indicateur reconnu à cette fin est l'*alpha de Krippendorff*.

2.1.1.2.2. Fiabilité

Le questionnaire est fiable s'il est stable, c'est-à-dire si les réponses ne dépendent ni de l'enquêteur, ni du moment auquel le questionnaire est administré. Il y aurait évidemment problème si les réponses devaient dépendre de qui est l'enquêteur, par exemple si un questionnaire d'évaluation des enseignements par les étudiants devait engendrer des réponses différentes suivant que l'enquêteur est ou non l'enseignant lui- même. De même, un questionnaire n'est pas fiable si les réponses diffèrent suivant le moment auquel il est administré, par exemple, dans le cas précédent, juste avant ou juste après l'examen terminal.

Pour déceler un problème de fiabilité, on peut administrer le questionnaire à deux groupes distincts detesteurs, suivant la méthode du *split-half*. Chaque groupe doit être d'effectif suffisant (au moins trente participants). Les groupes doivent être définis soit comme différents, soit comme identiques. Des critères de différence ou similarité doivent donc être préalablement établis, par exemple l'âge, la profession, le genre, le niveau d'études, les caractéristiques psychométriques, etc. On compare ensuite les réponses

Les différences entre questionnaires initial et parallèle doivent être de nature strictement formelle, touchant éventuellement à la l'introduction/présentation, mais surtout aux questions : vocabulaire utilisé (par exemple langue courante dans le questionnaire initial et langagetechnique dans le questionnaire parallèle), style (plutôt parlé/familier ou plutôt littéraire/formel), ton (plutôt direct ou plutôt distancié), canal d'administration (par exemple d'abord électronique puis ensuite en répondant

oralement à un enquêteur en face à face), ordre des questions, nombre de questions (par exemple une question dans le questionnaire initial est décomposée en deux questions dans le questionnaire parallèle), échelle d'évaluation (par exemple de 0 à 20 puis de 0 à 100, ou encore en faisant passer une échellede Likert de 5 à 7 modalités), etc. Plus les sources de différenciation sont nombreuses, plus le risque d'effet d'apprentissage est réduit.

Le questionnaire parallèle peut être considéré comme équivalent au questionnaire initial s'il engendre, pour chaque question ou modalité de réponse, les mêmes résultats auprès du même groupe de testeurs. Par exemple, le pourcentage de répondants choisissant tellemodalité de réponse à telle question est le même avec les deux questionnaires. Ou encore, si la réponse attendue est de nature quantitative, la moyenne des réponses est similaire dans les deux questionnaires. On ne peut évidemment s'attendre à une similarité parfaite, de sorte qu'un intervalle d'écart admissible doit être défini. Il n'existe pas de pratique consensuelle surl'écart qui peut être considéré comme admissible, doncla décision sur ce point est de la responsabilité du chercheur. Plusieurs méthodes de définition de l'écart admissible sont envisageables. Par exemple, on peut considérer que le pourcentage de répondants choisissant telle modalité de réponse à telle question nedoit pas varier de plus de 5 points (ou 10 points, ou 15 points par exemple) d'un questionnaire à l'autre. Si les réponses sont quantitatives, on peut considérer que le coefficient de corrélation entre les réponses à telle question du questionnaire initial et les réponses à la question équivalente du questionnaire parallèle doitêtre au moins de 85% (ou 90%, ou 95%, par exemple) ;ou encore que la différence de moyennes entre les deuxquestionnaires pour cette question ne doit pas être statistiquement significative, ou qu'elle doit être inférieure à un seuil. Mais quelle que soit la méthode choisie, le choix doit être argumenté, par exemple par référence à la littérature, ou à une théorie, ou aux spécificités de l'objet de recherche, etc. La règle est d'éviter l'arbitraire.

2.1.1.2.3. Cohérence interne

Le questionnaire est cohérent si les questions regroupées au sein d'un même domaine portent effectivement sur ce domaine. Si le questionnaire est cohérent, il n'y a dans

chaque domaine que des questions relatives à ce domaine. Une procédure courante pour vérifier la cohérence interne de questionnaires consiste à calculer le coefficient *alpha de Cronbach*[26] pour les différents domaines duquestionnaire. Dans la pratique, il est d'usage de considérer le domaine comme cohérent si l'alpha de Cronbach est supérieur à 0,7. À défaut, il importe d'identifier les raisons du niveau insuffisant de cohérence des domaines concernés et de remédier à cette insuffisance, par exemple en reformulant, ou en remplaçant, ou en retirant les items à l'origine du problème.

Par exemple, imaginons une recherche sur le climat scolaire qui, dans son questionnaire, considère le domaine « bien-être de l'élève à l'école », entre autres.

Le domaine comprend trois items. Le répondant notechaque item sur une échelle de 1-« très mauvais » à 5-
« très bien ». Les réponses obtenues auprès de quaranterépondants s'établissent comme suit :

Scores de 40 répondants à trois items du domaine « bien-être de l'élève à l'école »

Identifiant du répondant	Conditions matérielles de travail	Relations avec les autres élèves	Sentiment d'être bien encadré
321	2	4	2
322	3	4	3
323	1	4	2
324	3	5	4
325	5	4	4
326	3	2	3
327	4	3	4
328	2	3	5
329	1	2	4
330	5	5	1

331	5	5	1
332	4	1	1
333	3	5	2
334	3	4	4
335	2	3	1
336	5	4	1
337	4	4	3

2.1.2. Échantillonnage

Une fois au point, le questionnaire peut être administré,mais encore faut-il déterminer à qui. Il ne s'agit plus à ce stade de définir la population d'intérêt, mais dedéterminer lesquels de ses membres interroger. Parfois,la population d'intérêt est de taille limitée. C'est le caspar exemple dans une recherche sur les facteurs locaux de l'échec au bac dans une petite ville de quelques milliers d'habitants. Lorsque l'effectif de la population d'intérêt se limite à quelques dizaines ou centaines d'individus, on peut envisager d'administrer le questionnaire à l'ensemble de ces individus. Souvent cependant, l'effectif de la population d'intérêt peut être très élevé, atteignant par exemple plusieurs centaines de milliers d'individus. Dans ce cas, administrer le questionnaire à l'ensemble de ces individus peut soulever de redoutables problèmes de moyens (notamment trouver et financer des enquêteurs) et des problèmes d'ordre technique (en particulier la gestion et le traitement des données) de nature à rendre l'ensemble des processus de collecte et de traitement lourds, complexes, couteux et longs. La solution alternative consiste alors à travailler sur échantillon. Mais dans ce cas, l'objectif restant bien d'aboutir à des conclusions généralisables, il faut éviter l'échantillonnage sauvage, et au contraire procéder avec méthode. Deux principaux points méritent alors attention : la composition de l'échantillon, et sa taille.

2.1.2.1. Composition de l'échantillon

On distingue deux approches en matière de composition de l'échantillon : l'échantillonnage non- probabiliste et l'échantillonnage probabiliste.

2.1.2.1.1. Échantillonnage non-probabiliste

L'échantillonnage est non-probabiliste s'il est conduit sans prendre en compte la probabilité qu'a chaque membre de la population-mère d'être sélectionné pour faire partie de l'échantillon (« probabilité de sélection », on dit encore « probabilité d'inclusion »). Une méthode connue d'échantillonnage non-probabiliste est le micro-trottoir, qui consiste à se poster à un endroit et à interroger tout ou partie des passants. Il est clair que toute la population n'avait pas la même probabilité de passer à cet endroit à ce moment, donc que les probabilités de sélection des individus ne sont pas égales, (par exemple la probabilité de sélection des riverains est plus élevée), sans que l'on en sache plus sur le niveau exact de ces probabilités.

La méthode la plus courante d'échantillonnage non- probabiliste est la *méthode des quotas*. Le principe en est de constituer un échantillon tel que chaque composante de la population-mère soit représentée au sein de l'échantillon dans les mêmes proportions que dans la population-mère. Cette méthode repose sur le postulat que le phénomène que l'on cherche à étudier (par exemple l'absentéisme scolaire, ou encore tel ou tel style d'apprentissage, ou quoi que ce soit d'autre), même si on en ignore l'ampleur au sein de la population-mère, existera dans l'échantillon de la même façon qu'il existe au sein de la population-mère, dès lors que la structure connue de la population-mère est respectée. En d'autres termes, on suppose qu'il existe une certaine corrélation entre toutes les caractéristiques possibles d'une même population.

D'un point de vue théorique, la limite de la méthode des quotas est que son postulat fondateur reste à démontrer. D'un point de vue pratique, une de ses limites tient au fait qu'il n'existe pas de règle quant aux caractéristiques à prendre en compte pour définir la structure d'une population. Les caractéristiques possibles sont innombrables : dans l'absolu, elles peuvent être démographiques, sociales, éducationnelles, professionnelles, culturelles, religieuses, politiques, économiques, géographiques, par exemple. Sélectionner certaines caractéristiques et pas d'autres est donc toujours discutable, et doit par conséquent être argumenté. Cependant, en admettant même que

toutes les caractéristiques possibles puissentêtre prises en compte, cela impliquerait de prendre en compte de micro-composantes de la population, dont l'effectif est restreint à quelques unités ou dizaines d'unités. Il faudrait alors que ces micro-composantes soient représentées dans l'échantillon dans la même proportion que dans la population-mère. Par suite, il faudrait inclure dans l'échantillon suffisamment d'individus appartenant aux composantes numériquement les plus importantes de la population- mère pour que ces composantes importantes soient représentées à leur exacte proportion dans l'échantillon. Il en résulterait que l'effectif de l'échantillon risquerait de n'être plus très éloigné de celui de la population-mère, ce qui à l'évidence annulerait l'intérêt même de recourir à un échantillon.

2.1.2.1.2. Échantillonnage probabiliste

La forme canonique de l'échantillonnage probabiliste est l'*échantillonnage aléatoire simple*. Le principe de l'échantillonnage aléatoire simple est que les personnes auxquelles le questionnaire sera soumis sont sélectionnées au hasard, de sorte que toutes ces personnes ont chacune la même probabilité n/P d'être sélectionnées pour faire partie de l'échantillon (n étant la taille de l'échantillon et P l'effectif de la population-mère). La méthode nécessite de disposer de la liste complète des individus appartenant à la population- mère (« base de sondage »), et d'attribuer à chacun de ces individus un numéro d'ordre. On procède ensuite au tirage au sort en utilisant un générateur de nombres aléatoires.

Plusieurs variantes de l'échantillonnage probabiliste existent. L'échantillonnage aléatoire « systématique » (ou « échantillonnage par intervalles »), d'abord, dans lequel le premier membre de l'échantillon est tiré au sort, puis chaque n-ième membre de la base de sondage à partir du premier sélectionné est inclus dans l'échantillon jusqu'à ce que celui-ci soit au complet.

2.2. L'ENQUÊTE PAR ENTRETIEN

La littérature relative aux méthodes d'entretien est riche. De façon générale,

l’entretien se définit comme un échange verbal entre deux personnes, à l’initiative d’un chercheur et contrôlé par lui, en vue d’obtenir du répondant des informations pertinentes par rapport à unobjet d’étude.

On en distingue deux cas polaires. Le premier estl’entretien libre / non-directif, dans lequel le chercheur adapte librement l’ordre et la formulation des questions (à partir toutefois d’un « guide d’entretien » préétabli),tandis que le répondant est maitre du format de ses réponses. Dans le deuxième cas, l’entretien est structuré, directif et standardisé. Le chercheur (ou l’enquêteur qu’il a mandaté) ne s’écarte pas de l’ordre et de la formulation des questions fixés dans le guide d’entretien. Dans sa forme extrême, l’entretien standardisé ne laisse au répondant que la possibilité dechoisir ses réponses sur une liste qui lui est proposée. L’entretien standardisé permet de comparer terme à terme les réponses de répondants différents. Entre ces deux cas polaires, l’entretien semi-structuré (« semi- directif ») impose l'ordre et la formulation des questions, mais prévoit des questions ouvertes, qui laissent au répondant une certaine marge dans le formatde ses réponses.

Chapitre 3. Préparer les données

Les données destinées à une analyse quantitative sont en principe regroupées dans une feuille de calcul de tableur (par exemple Excel). Lorsque le questionnaire a été administré en ligne, les plateformes d'enquête permettent au chercheur de récupérer directement le fichier des données. Dans les autres cas, le répondant communique ses réponses oralement ou sur document papier à l'enquêteur / chercheur qui les saisit dans la feuille de calcul. Le principe général est d'organiser la feuille de calcul de façon à avoir une colonne par question ou par modalité de réponse, et une ligne par répondant. Les réponses sont saisies telles que fournies par le répondant lorsque la variable est quantitative ou textuelle courte (commentaire libre). Lorsque la variable est qualitative et les modalités de réponse possibles préétablies, il est généralement pluscommode de représenter chaque modalité de réponse par un numéro de code.

Une fois le fichier rempli, il doit être finalisé. La finalisation exige d'abord de scruter les données à la recherche d'erreurs à éliminer. Les sources de possible erreur sont multiples. Des erreurs peuvent intervenir aussi bien lorsque le répondant fournit ses réponses que lors de la saisie de ces dernières dans le fichier. Dans les questionnaires auto-administrés, certains répondants peuvent aussi parfois fournir des réponses fantaisistes.

Mais la finalisation renvoie aussi à deux autres points importants : le traitement des valeurs aberrantes et la standardisation des données.

3.1. VALEURS ABERRANTES

Même en l'absence d'erreurs, les données peuvent contenir des valeurs « aberrantes » (*outliers*), inhabituellement faibles ou élevées, et plus généralement non-représentatives en ce que leur profil tranche radicalement avec le ou les profils habituellement observés dans les données sur lesquelles porte l'analyse. Elles peuvent être détectées à la lecture des données, ou en visualisant ces dernières au moyen de graphiques (par exemple nuages de points, histogrammes ou courbes). Les logiciels statistiques proposent aussi souvent des tests de détection des valeurs aberrantes. Les

valeurs aberrantes peuvent sensiblement influencer la moyenne de la série ou le résultat de tel ou tel calcul, et engendrer une impressiond'ensemble incorrecte. On peut vérifier l'incidence des valeurs aberrantes en effectuant les calculs avec puis sans elles, et comparer les résultats. Il importe que le chercheur s'interroge, en fonction de la nature spécifique de sa recherche, sur la nécessité de retirer ou non les valeurs aberrantes, et plus généralement sur la façon de les gérer.

3.2. STANDARDISATION

Un traitement préparatoire additionnel qui peut s'avérer nécessaire dans la phase de finalisation du fichier des données est la standardisation des données. La problématique ici est celle de données d'échelles différentes qui, par conséquent, ne peuvent être visualisées sur un même plan, et par ailleurs ne sont pas directement comparables. C'est par exemple le cas lorsque les séries considérées sont exprimées en unitésde compte différentes (nombres de personnes, nombre d'incidents, unités de temps, salaire, etc.), ou sont d'ordres de grandeur disproportionnés, allant parexemple de 1 à 10 dans la première série et de 100 à 1000 dans la seconde.

Pour rendre ces données comparables, on procède habituellement à leur standardisation.

Il existe un grand nombre de méthodes de standardisation, adaptées à différents types de problèmes. On présente ici quelques formes destandardisation parmi les plus courantes : la standardisation par centration-réduction, les standardisations max-min, et la standardisation par moyenne ou écart-type.

3.2.1. Standardisations max-min

On distingue trois variantes de standardisation max- min : la standardisation max-min en intervalle [0 ; 1] ; la standardisation max-min en intervalle [-1 ; 1] ; et la standardisation max-min en intervalle quelconque.

3.2.2.1. Standardisation max-min en intervalle [0 ; 1]

La méthode consiste à transformer la série de sorte que toutes ses valeurs soient comprises dans un intervalle [0 ; 1]. Pour ce faire, on remplace chaque valeur x_i de la série par

$$x_i^* = \frac{x_i - Min}{Max - Min}$$

où

Min est le minimum de la série considérée ; et

Max son maximum.

La série sera par conséquent transformée en une série bornée par 0 (transformée du minimum de cette série) et 1 (transformée du maximum de la série). On voit qu'iln'y a pas, avec cette méthode, distribution des valeurs standardisées autour de zéro, comme c'était le cas avecla standardisation par centration-réduction.

Comme l'illustre par exemple le Graphique 3.3, lastandardisation des données d'une série par cette méthode ne modifie pas la structure de la courbe de cette série.

Deuxième Partie : Analyser les données

Une fois les données collectées et préparées, il s'agit de les analyser. La gamme des méthodes d'analyse quantitative de données est riche et en constante évolution. La première étape pour le chercheur consiste à sélectionner la méthode pertinente à mettre en œuvre pour sa recherche. Le critère le plus important à prendre en compte pour effectuer cette sélection est l'objectif de l'analyse. Par exemple, les méthodes descriptives peuvent être suffisantes si l'objectif est de décrire un échantillon sans intention d'en inférer des conclusions généralisables à l'ensemble de sa population-mère. Il est donc important d'être, dès le départ, au clair avec l'objectif de l'analyse, afin de sélectionner la méthode la plus appropriée, et d'être ensuite en mesure de justifier son choix.

Il n'est pas rare cependant que pour un objectif donné, plusieurs méthodes concurrentes soient disponibles. Par exemple, on peut comparer des moyennes en appliquant des tests d'hypothèses ou en effectuant une analyse de variance. Lorsque plusieurs méthodes sont disponibles au service d'un même objectif, d'autres critères de choix peuvent intervenir, par exemple le fait que certaines méthodes peuvent être plus lourdes à mettre en œuvre que d'autres. De façon générale, il est essentiel d'éviter l'arbitraire, d'examiner rigoureusement toutes les facettes du problème, d'effectuer un choix de méthode, et de soigneusement étayer son choix. En revanche, la pratique de la « triangulation », fréquente en recherche qualitative et qui consiste à approcher le même problème par plusieurs méthodes concurremment, n'est pas vraiment d'usage courant en recherche quantitative.

Enfin, toutes les méthodes d'analyse quantitative nécessitent que des conditions de validité soient remplies. La vérification de ces conditions fait partie du travail d'analyse et contribue à déterminer la validité et la fiabilité des résultats obtenus. Le cas échéant, la phase de vérification des conditions de validité permet de repérer et de traiter des sources de fragilité. Rendre compte de la vérification des conditions de validité fait partie intégrante du rapport d'analyse. Les conditions qui s'avèrent ne pas être

remplies doivent êtreexplicitement signalées dans le rapport d’analyse comme autant de réserves sur la validité des résultats obtenus.

Chapitre 4. Lexicométrie : l'étudequantitative de textes

L'application des méthodes quantitatives n'est pas réservée aux données numériques. Les données textuelles peuvent, elles-aussi, être traitées de façon quantitative. Il en va ainsi des réponses aux entretiens et des réponses aux questions ouvertes de questionnaires, et plus généralement de tout texte littéraire, journalistique ou politique. Habituellement, les données textuelles sont traitées par analyse de contenu, c'est-à-dire un ensemble de méthodes qualitatives qui permettent d'identifier la signification d'un discours à partir de ses caractéristiques lexicales et syntaxiques, et de cerner l'univers de référence et les attitudes du locuteur (Bardin, 2013). La lexicométrie vise, elle aussi, à appréhender le sens d'un propos, mais privilégie les observations quantitatives. Les deux types d'approches ont des points communs : l'analyse de contenu fait elle aussi appel au dénombrement d'occurrences, tandis que la lexicométrie de son côté nécessite l'identification de catégories sémantiques.

4.1. ÉTABLIR LE CADRE DE RÉFÉRENCE

La première étape consiste à établir l'univers conceptuel et théorique dans le cadre duquel le matériau textuel (*corpus*) sera analysé. Il s'agit ici de répondre à la question : quelle est la théorie et quels sont les concepts à la lumière desquels l'objet d'étude peut être analysé. Le cadre de référence indique ce quedevrait être, d'après la théorie, le contenu sémantique du corpus étudié : quels thèmes, quelles idées, quelles catégories, quels termes. Les catégories elles-mêmes sont de différents types (catégories d'acteurs, de contextes, d'actions, de moyens, etc.).

En découle une *grille d'analyse* du contenu du corpus.

Imaginons par exemple une recherche sur la gouvernance de l'école. Il est prévu d'effectuer une enquête par entretien auprès d'acteurs de terrain. Les répondants seront interrogés sur les défis prioritaires à relever au cours de la prochaine décennie. Une grille pour guider l'analyse des réponses qui seront obtenuespourrait comporter

les éléments suivants :

- catégories d'acteurs possibles : gouvernement, autorités publiques régionales et locales, directions d'établissements, représentants des personnels, représentants des élèves, représentants des familles, représentants des milieux économiques ;

- processus de décision : cogestion, concertation, votes, consultation facultative, consultation obligatoire, avisconforme, autocratie ;
- processus d'interaction : échange d'informations, négociations, conflits sociaux, arbitrages, procédures judiciaires ;
- processus de régulation : inspections, audits externes, contrôles administratifs, contrôles hiérarchiques, évaluation par les pairs, auto-évaluation, évaluation par les usagers ;
- outils d'élaboration du projet collectif : prospective, expérimentations, recherche académique,planification, innovation, réforme ;
- valeurs partagées : transparence, participation, équité, égalité, fiabilité, hiérarchie, respect, liberté, compétence, éthique, humanisme, solidarité, efficacité, légalité, développement professionnel, internationalisation, autonomie, responsabilité, intégration sociale, excellence.

4.2. RÉPERTORIER LES FORMES PAR FRÉQUENCED'OCCURRENCE

Il s'agit ici d'établir la liste des termes / mots (« formes ») contenus dans le corpus, classés par fréquence d'apparition. L'analyse se concentre sur les formes ayant un contenu sémantique, et ignore articles, pronoms, prépositions, conjonctions, et tous autres éléments de structure grammaticale, sauf bien entendu si le style littéraire fait partie de l'objet d'étude.

Les synonymes, d'abord, doivent être identifiés. Ils expriment en effet un même contenu sémantique, or ils apparaissent dans la liste primaire comme des formes distinctes, donc comptabilisées séparément. Il importe donc d'établir des listes de synonymes, de façon à disposer d'une vision plus exacte de la récurrence d'uncontenu sémantique dans le corpus.

Les associations de formes sont des expressions dans lesquelles deux formes sémantiques ou

plus sont utilisées conjointement. Or une forme n'a pas forcement la même signification lorsqu'elle est utiliséeseule ou dans une expression. Il importe donc de pouvoir repérer les associations de formes, de façon à distinguer leur signification spécifique de celle desmêmes formes employées isolément. La signification des associations de formes (ou des formes elles-mêmesdu reste) peut être notamment précisée grâce aux
« concordances », c'est-à-dire aux contextes / phrases dans lesquelles ces associations apparaissent. Chaque association de formes ayant une signification spécifique doit être répertoriée sur la liste des formes, et ses occurrences faire l'objet d'une comptabilisation distincte.

4.3. ANALYSE SÉMANTIQUE

Il s'agit ici de mettre en regard la fréquence observée des différentes formes avec la grille de référence établieen étape 1. L'analyse consiste à identifier et commenter, par référence à la grille :

- les idées, catégories, thèmes et termes qui figurentdans le corpus
  - Signification
  - Hiérarchie dans l'insistance
- les idées, catégories, thèmes et termes qui n'y figurentpas ;
- les idées, catégories, thèmes et termes qui constituentdes originalités/innovations ;
- les représentations, explicites ou implicites ;
- les attitudes et positionnements ;
- les non-dits et arrières pensées ;
- les idéologies sous-jacentes ;
- les incohérences ; etc.

L'analyse s'effectue pour l'ensemble du corpus. Elle peut s'effectuer aussi, en outre, pour chaque partie du corpus, de façon à mettre en évidence l'éventuellespécificité de telle ou telle partie du corpus, par exemple en termes de sur- ou sous-représentation de telou tel contenu sémantique par date de discours ou (typede) répondant.

Chapitre 5. Analyse classificatoire : discernerdes groupes homogènes

L'analyse classificatoire s'utilise dans le cas où on cherche à répartir un ensemble de données en sous- groupes homogènes. On dispose d'observations relatives à un ensemble d'individus, et on cherche à répartir ces individus en groupes dont les membres sontplus proches les uns des autres qu'ils ne le sont du restede la population. C'est le cas par exemple lorsqu'on cherche à distinguer des profils, des groupes de comportements, des groupes de performances, etc.

La problématique méthodologique de l'analyse classificatoire est différente de celle du calcul de quartiles, déciles ou centiles. En premier lieu, lorsqu'on décompose une série en quarts, dixièmes ou centièmes, l'affectation d'un individu à un intervalle s'effectue au regard d'une unique variable. Considérons par exemple les stagiaires inscrits dans un établissement de formation, on pourra bâtir des intervalles interquartilesau regard du score au test d'entrée, ou au regard du nombre d'heures de formation requises pour une remise à niveau, ou au regard de l'âge, etc. L'analyse classificatoire, elle, peut permettre de constituer des groupes homogènes au regard de plusieurs variables simultanément, réunissant au sein d'un même groupe des individus proches en termes à la fois d'âge, de besoin de formation et de performance au test d'entrée.

En deuxième lieu, dans le découpage en intervalles, les classes / intervalles sont, par définition, d'effectifs identiques. Dans l'analyse classificatoire, au contraire, la taille des groupes n'est pas prédéterminée. La taille d'un groupe dépend uniquement du nombre d'individussuffisamment proches pour pouvoir faire partie de ce groupe.

En troisième lieu, dans un intervalle, un individu situé à la borne supérieure est plus proche de l'individu situé à la borne inferieure de l'intervalle suivant qu'il ne l'est de l'individu situé à l'autre borne de l'intervalle auquel lui-même appartient. L'analyse classificatoire, au contraire, permet de n'inclure dans le groupe que les individus les plus proches entre eux au regard de toutesles variables d'intérêt.

5.1. PARTITIONNEMENT UNIVARIÉ

La méthode du partitionnement univarié permet de traiter les cas dans lesquels on dispose de données sur une unique variable quantitative (X) pour plusieurs individus i $= 1, \cdots, n$, et où on cherche à répartir ces individus en groupes homogènes au regard de cette variable. Le partitionnement univarié permet d'effectuer cette classification. Le nombre de classes doit être défini par le chercheur, en fonction de la théorie dans le cadre de laquelle l'analyse s'inscrit.

5.2. MÉTHODE K-MEANS

La méthode k-means (ou méthode des nuées dynamiques) permet de traiter les cas dans lesquels on dispose d'observations relatives à plusieurs variables quantitatives (X_1, X_2, , $\cdots$, $X_£$) pour plusieurs individus (i $= 1, 2, \cdots, n$). On cherche à répartir ces individus en groupes homogènes au regard de l'ensemble de ces variables. La méthode permet d'effectuer cette classification. Le nombre de classes doit être défini par le chercheur, en fonction de la théorie dans le cadre de laquelle l'analyse s'inscrit.

5.3. CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

Comme dans la méthode k-means, on dispose d'observations relatives à plusieurs variables quantitatives (X_1, X_2, , $\cdots$, $X_£$) pour plusieurs individus (i $= 1, 2, \cdots, n$). On cherche à répartir ces individus en groupes homogènes au regard de l'ensemble de ces variables. Mais à la différence de la méthode k-means, la classification ascendante hiérarchique (CAH) laisse au chercheur le choix de fixer ou non *a priori* le nombre de groupes. La méthode permet en effet, en l'absence de contrainte théorique, de déterminer par calcul le nombre de groupes qui permet d'assurer la plus grande homogénéité / proximité intra-groupe.

5.4. CLASSIFICATION EN CLASSES LATENTES

La classification en classes latentes permet de traiter les cas dans lesquels on dispose d'observations relatives à plusieurs variables qualitatives pour plusieurs individus. S'agissant de variables qualitatives, on ne peut calculer de distance entre individus. L'alternative est alors de calculer les covariations entre variables. La méthode repose sur le postulat que les variables qualitatives observées sont en fait liées à des variables qualitatives sous-jacentes, non directement observables, mais qui les déterminent. Par suite, les variables qualitatives observées qui varient de la même

façon sont présumées appartenir à une même modalité d'une variable qualitative nominale sous-jacente (classe latente). Le chercheur peut, en fonction de son cadre théorique, fixer le nombre de classes latentes, ou identifier le nombre optimal au regard d'une batterie decritères.