Introduction aux statistiques sociales.

Une approche intuitive

El Hadj Touré, Ph.D.

elhadj.toure@umontreal.ca

Manuel en cours de rédaction

Soumission aux Presses de l'Université de Montréal

(Décembre 2023)

Chapitre 1.
Cheminement en recherche quantitative

Les statistiques sont intimement liées aux questions de méthodologie en sciences sociales. D'une part, les principes généraux qui guident la recherche orientent en même temps l'analyse statistique et déterminent la qualité des conclusions qui découlent de l'interprétation des résultats. D'autre part, le recours aux statistiques amène à exprimer en des termes quantitatifs le problème de recherche et la façon d'y répondre en interrogeant le réel. Tout naturellement, dans ce manuel de statistiques sociales, nous commençons par un chapitre introductif qui documente les questions de méthodologie quantitative. Certes la maîtrise des statistiques s'avère indispensable pour les sciences sociales. Mais, elle n'est utile que si l'on est à même de mener une étude générant des données probantes exploitables par des outils statistiques appropriés. Considérant l'aspect technique de l'ouvrage, ce chapitre fait un survol des différentes étapes de la méthodologie quantitative, un préliminaire essentiel mais trop souvent négligé. L'objectif est d'amener le lecteur à mieux situer la place des statistiques dans le processus de quantification.

A la fin de ce chapitre, le lecteur sera en mesure de cheminer convenablement dans une recherche quantitative. Précisément, il devra être capable de :

1. Expliquer clairement, par des exemples concrets, ce qu'est une recherche quantitative et ce qu'elle n'est pas ;
2. Définir une problématique de recherche sous l'angle de l'approche quantitative, tout en étant en mesure de passer de la question générale à la question spécifique ;
3. Formuler une hypothèse de recherche et la traduire en des termes qui la rendent opérationnelle et intelligible au raisonnement statistique ;
4. Savoir collecter des données numériques, primaires et secondaires, tout en étant en mesure de différencier le devis expérimental et l'enquête par questionnaire ;
5. Avoir un aperçu des techniques statistiques de base, descriptives et inférentielles, couramment utilisées dans l'analyse des données numériques ;
6. Savoir interpréter les résultats qui en résultent pour leur donner sens et surtout différencier l'interprétation statistique de l'interprétation sociologique/théorique ;

Au niveau de chacune de ces étapes, la maîtrise des statistiques est essentielle pour mener convenablement une recherche quantitative. Même si les caractéristiques de la recherche quantitative sont mises en avant, elles sont parfois comparées à celles de la recherche qualitative. Toutefois, il ne s'agit pas d'opposer ces deux approches, qui contribuent chacune à sa manière et parfois ensemble, à l'acquisition de connaissances empiriques à propos de phénomènes étudiés en sciences sociales. La comparaison vise plutôt à spécifier et mieux accéder au raisonnement statistique sous-jacent à la démarche quantitative.

NE PAS COPIER

1.1. Qu'est-ce que la recherche quantitative ?

En sciences sociales, on mène souvent une recherche pour acquérir des connaissances scientifiques de nature empirique. À cet effet, il existe deux grandes familles de méthodes de recherche : les méthodes quantitatives et les méthodes qualitatives.

Généralement, lorsqu'on parle de recherche quantitative, on pense à certains termes clés : nombres, statistiques, quantité, données, fait, fréquence, explication, causalité. Lorsque l'on amalgame tous ces termes, on peut avoir une idée assez claire de ce qu'est la recherche quantitative. En fait, elle tente de *décrire* et d'*expliquer* un phénomène *quantifiable* en s'appuyant sur la manipulation de *données numériques* issues de l'observation de la réalité.

L'activité de quantifier est au fondement de la plupart des sciences sociales, y compris la sociologie, la psychologie, la démographie, l'économie, la science politique et les sciences de l'éducation. Elle réfère au fait d'exprimer sous forme numérique des faits, comportements, opinions et attitudes à propos de phénomènes soumis à une étude scientifique. Une fois que les données sont obtenues et exprimées en nombres, il faut les traduire en informations chiffrées. C'est là qu'interviennent les statistiques. Celles-ci permettent de rendre informatives les données numériques sur la base d'opérations mathématiques. Par exemple, on sait qu'en 2020, 17% des Québécois âgés d'au moins 12 ans fument du tabac. Ce pourcentage est une statistique qui décrit l'ampleur du tabagisme au Québec, phénomène qui peut être expliqué entre autres par l'environnement familial.

On se passerait difficilement des statistiques. Permettez-moi encore d'être plus clair : sans statistiques aucune recherche quantitative n'est envisageable. Elles aident non seulement à supporter ou éclairer une décision, mais surtout à répondre à des questions qu'on se pose à propos d'un phénomène, et ce de façon précise et objective. Les statistiques, en réalité, ont la particularité d'être précises et objectives. *Précises*, puisqu'elles ne comportent aucune ambiguïté. *Objectives*, en ce sens qu'elles font l'objet d'un consensus relativement. Par exemple, lorsqu'on affirme que les Québécois sont 17% à fumer du tabac comparativement à 15% des autres Canadiens, l'affirmation repose sur des faits précis et objectifs.

Mais attention, précision ne signifie pas exactitude. La tentation est grande de croire que les statistiques produisent systématiquement des informations exactes. Or, l'analyste est susceptible de se tromper dans l'opération mathématique lors du calcul des pourcentages et même dans la mesure du phénomène étudié. Il peut poser la question de fumer ou non de telle sorte qu'elle soit comprise différemment par les répondants.

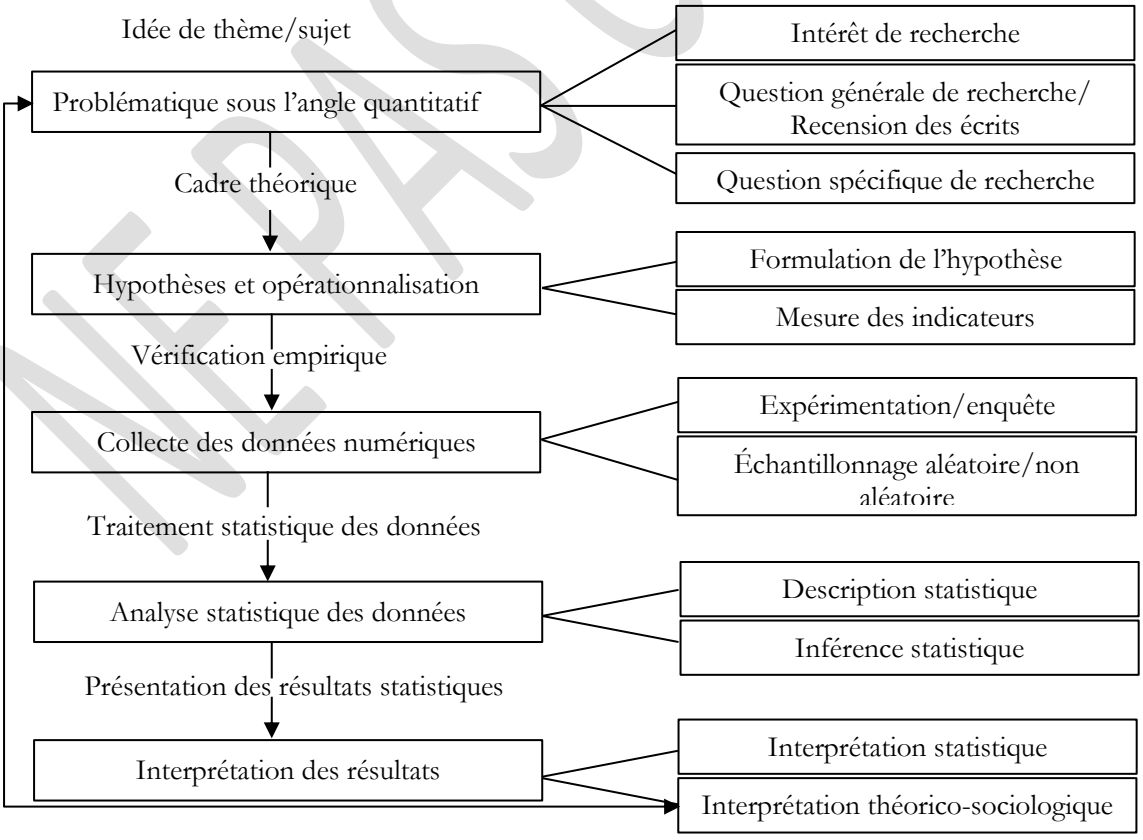
De même, objectivité ne signifie pas neutralité. Les chiffres relèvent d'une construction du chercheur. Ils ne sont pas donnés naturellement. Leur production s'opère par le biais de données dont la collecte peut être orientée dans un sens ou un autre. Par exemple, une fabrique de tabac peut commanditer un sondage ou elle sous-estime le nombre réel de fumeurs atteints de cancer. Au même moment, un organisme de défense des droits humains surestimerait cette même proportion. Si objectives soient-elles, les statistiques ne sont pas neutres. Elles peuvent être instrumentalisées à outrance selon les acteurs en présence et les intérêts du moment. Il en est ainsi lorsque le docteur Mailloux, au Québec, s'appuyait sur des statistiques pour soutenir l'hypothèse selon laquelle les Blancs étaient plus intelligents que les Noirs. Nous verrons que la relation entre l'ethnie et le quotient intellectuel relève

de ce qu'on appelle, dans le jargon statistique, une *relation fallacieuse*. Pour chaque technique abordée dans cet ouvrage, nous mettrons en évidence comment elle peut faire l'objet d'une instrumentalisation à des fins inavouées, parfois voulues, parfois non voulues.

Au demeurant, le risque est grand de réduire la recherche quantitative aux statistiques. Il n'en est rien. Des idées viennent en amont des statistiques. Elles nous suggèrent un sujet à étudier, des questions à soulever, des hypothèses à vérifier, des variables à mesurer, des relations à établir et des données à recueillir. Également, des idées viennent en aval des statistiques. Elles nous amènent à interpréter les résultats statistiques de façon à leur conférer un sens sociologique et à les inscrire dans un raisonnement théorique convaincant, alimentant ainsi de nouvelles recherches.

Contrairement à ce que l'on pourrait penser, les statistiques ne se réduisent pas non plus à des chiffres magiques, ni à des opérations mathématiques sur des nombres. Leur usage requiert une construction de l'esprit, de l'*imagination sociologique*, pour emprunter l'expression du sociologue Mills (1959). Aussi, comme interface entre données et interprétation, les statistiques relèvent-elles des méthodes quantitatives. Celles-ci réfèrent au processus de quantification à l'œuvre dans une démarche scientifique. Ce processus consiste à mettre en chiffre le monde social afin de mieux comprendre son fonctionnement. La figure ci-dessous illustre sous forme schématique le cycle et les différentes étapes d'une recherche quantitative. Nous survolons à présent chacune de ces étapes interreliées.

Figure 1.1. Le cycle et les étapes d'une recherche quantitative



NE PAS COPIER

1.2. Construction de la problématique sous l'angle quantitatif

Construire une problématique sous l'angle quantitatif revient à formuler une question de recherche opérationnelle dont la réponse peut être obtenue au moyen de la mobilisation de techniques statistiques appropriées. Précisément, le chercheur doit savoir passer de la question générale (1.2.2) à la question spécifique (1.2.3), tout en les élaborant de façon à les rendre intelligibles au raisonnement statistique. Auparavant, abordons quelques éléments essentiels qui permettent de définir l'intérêt de recherche (1.2.1).

1.2.1. L'intérêt de recherche

Une recherche débute toujours par le choix d'un thème d'intérêt. Si dans la plupart des situations de recherche, on part d'une idée relativement vague (thème), dans d'autres on a au départ une idée assez claire ou documentée du sujet à étudier. Dans les sciences sociales en général, le thème se rapporte à des phénomènes comme le suicide, l'itinérance, la criminalité, la santé mentale, les inégalités salariales, l'absentéisme ou le décrochage scolaire, la délinquance, la préférence politique, les technologies de l'information et de la communication, et même la quantification ou l'anxiété statistique. Quel que soit le type de recherche, quantitatif ou qualitatif, le chercheur peut s'intéresser à n'importe quel domaine du comportement humain. De même, n'importe quel niveau d'investigation peut être privilégié : macro (structures sociales), méso (organisations, groupes) ou micro (individus). Mais, comment né l'intérêt de recherche ? Construire une problématique consiste avant tout à trouver des idées de thème/sujet sur le monde social et son fonctionnement. Le sociologue Fox (1999) identifie au moins quatre sources de provenance des idées. Nous les empruntons et adaptons dans cet ouvrage.

Tout d'abord, le chercheur peut se référer aux *théories* à l'œuvre en sciences sociales. Parce qu'elles s'insèrent dans des paradigmes incarnant une certaine vision de la société, les théories peuvent orienter les programmes ou projets de recherche. Elles proposent en effet des problématiques à privilégier et des façons d'y répondre partagées par une communauté ou des équipes de chercheurs. Nous savons également qu'une théorie est un ensemble cohérent de concepts qui aident à expliquer, prédire des comportements humains à l'aide de propositions logiquement liées et empiriquement vérifiables. Ces concepts peuvent amener les chercheurs à s'intéresser à certaines thématiques plutôt que d'autres. Par exemple, en sociologie, le fonctionnalisme postule que les institutions sociales concourent au maintien de l'équilibre social. Cette théorie peut donc susciter des idées intéressantes sur les fonctions de la famille et leur évolution actuelle. Si le chercheur se réfère plutôt à l'analyse stratégique, peut-être s'intéressera-t-il aux stratégies développées par les acteurs pour s'affranchir du caractère contraignant des systèmes organisationnels. Il peut selon cette perspective étudier l'absentéisme scolaire ou le syndicalisme.

Le choix d'un thème/sujet de recherche peut ensuite être guidé par toutes sortes de lectures qui fournissent des informations à propos des phénomènes sociaux. Que ces lectures portent sur des travaux scientifiques, des publications de divers organismes, des journaux, magazines, romans ou essais. Par exemple, la lecture d'un article sur l'absentéisme scolaire en France peut suggérer la réplication de l'étude à d'autres contextes. C'est cela qui m'avait inspiré dans le cadre d'un dossier de recherche effectué en première année de sociologie.

J'y mettais en relation la lourdeur du système d'enseignement et l'absentéisme des étudiants au sein de la Section de sociologie de l'Université Gaston Berger de Saint-Louis (Sénégal).

Le chercheur peut aussi vouloir fournir des réponses à des problèmes sociaux pratiques qui le préoccupent en tant qu'individu vivant dans la société. Certes la quête de l'objectivité est un trait caractéristique de la recherche quantitative. Malgré la prédominance de cette posture positiviste, le sociologue ne peut pas s'exclure complètement de sa vision de la société, de ses propres valeurs ou de celles de sa communauté. Les enjeux sociétaux sont susceptibles de générer en lui des idées de thème de recherche. C'est l'exemple d'un étudiant qui œuvre pour améliorer le sort des immigrants en investiguant les inégalités liées à l'accès au travail au Canada. De même, un activiste sensible aux questions environnementales et désireux de changer les mentalités peut s'intéresser à la croyance face au changement climatique. Vous êtes engagés en faveur d'un enjeu, de votre société ou communauté ? N'hésitez pas à mener des recherches qui contribuent à changer les choses.

Finalement, on peut s'inspirer d'observations simples du monde qui nous entoure. Le sociologue, comme tout spécialiste des sciences sociales, est un observateur perpétuel de son environnement sociétal. Le sens commun observe aussi des phénomènes de mode comme les jeans troués, le piercing ou le tatouage chez les jeunes. Evidemment, cette forme de « sociologie savante » ne peut pas être une source de connaissance crédible. On s'imagine facilement connaitre la société, de telle sorte qu'on reproche parfois aux sociologues de rechercher ce que les gens savent déjà. On entreprend justement une recherche scientifique pour tenter de décrire et d'expliquer, avec rigueur et méthode, les phénomènes de mode observés dans la vie quotidienne[1]. Mais d'emblée, l'observation de ces phénomènes peut suggérer des idées concernant le développement de sous-cultures chez les jeunes.

Peu importe son origine, le choix d'un thème/sujet de recherche est intimement lié à l'intérêt qu'il suscite chez le chercheur. Commandité, proposé ou pas, il peut s'agir d'une pertinence sociale ou scientifique. En effet, l'intérêt pour un phénomène peut se justifier par sa nouveauté, sa rareté, son ampleur, ses conséquences importantes sur la société, la nécessité de le connaître davantage ou l'urgence d'élucider le problème qu'il contient, le besoin de tester des théories ou instruments de mesure, etc. Une fois le thème/sujet de recherche choisi et son intérêt explicité, il faut le problématiser, c'est-à-dire le traduire au préalable en un questionnement qui s'inscrit dans une quête de la connaissance.

1.2.2. De la question générale de recherche…

Au départ d'une recherche sociale, il y a un thème ou un sujet qui suscite un problème social. Que ce problème soit explicite ou implicite. Un *problème social* est un phénomène, un comportement, une attitude qui pose problème à la société en raison de son ampleur et de ses conséquences importantes sur celle-ci. Par exemple, la crise sanitaire de la Covid-19 implique d'énormes coûts pour les sociétés. Mieux, nonobstant son caractère répandu, les

[1] À ce propos, dans les *Règles de la méthode sociologique*, Durkheim (1894) affirmait ceci : « C'est que, en effet, la réflexion est antérieure à la science qui ne fait que s'en servir avec plus de méthode ». La méthode repose sur la formulation d'hypothèses susceptibles d'être réfutées par des expériences reproductibles.

NE PAS COPIER

gens ne sont pas tous égaux devant cette pandémie. Également, lorsqu'il étudia le suicide, Durkheim (1897)[2] considérait ce phénomène comme une pathologie éminente des sociétés modernes. Il révéla la permanence de ce phénomène à travers le temps et sa variabilité d'une société à l'autre, chaque société étant prédisposée à fournir un nombre déterminé de morts volontaires. Lorsqu'on cherche à aller au-delà d'une simple constatation pour s'efforcer notamment de décrire ou d'expliquer un phénomène, on génère un problème sociologique. Un *problème sociologique* s'apparente donc à un problème de recherche. Il reflète un contraste apparent entre une réalité sociale et une situation désirable. Durkheim désirait expliquer la variabilité du suicide, car on ne connaissait pas à l'époque les causes véritables de cette pathologie. Un chercheur peut souhaiter décrire les caractéristiques des personnes décédées de la Covid-19, car il s'agit d'une pandémie nouvelle encore mal connue. Autant de problèmes généraux qui peuvent guider au départ des recherches sociales.

Pour définir sa problématique, le chercheur commence par formuler le problème de façon générale, soit l'écart entre ce qui est réel et ce qui est désirable. Comme le nom le laisse penser, un problème général de recherche est certes plus ou moins vague. Mais, il délimite déjà le thème puisqu'il inclut un choix parmi les multiples aspects du phénomène étudié. Le problème général de recherche peut être exprimé sous forme de question pour mieux indiquer l'intention de combler l'écart entre un constat dans la réalité et une attente à satisfaire. On parle alors de *question générale de recherche*. Quelles sont les causes du suicide ? Quelles sont les caractéristiques des personnes décédées de la Covid-19 ? En recherche quantitative, la question générale consiste plus souvent qu'autrement à vouloir décrire un phénomène et ses différentes composantes (prévalence, représentativité, variabilité, comparaison ou classification des individus) ou expliquer un phénomène en établissant des relations plausibles (association, prédiction, causalité). Ce faisant, le chercheur clarifie le *but* de la recherche en des termes qui justifient l'usage des méthodes statistiques.

La question générale de recherche constitue la deuxième phase de la construction d'une problématique, après celle de la définition de l'intérêt de recherche. Elle est encore appelée *question de départ*. Celle-ci est tellement générale qu'il s'avère difficile d'y répondre en une seule étude. Il faut donc aller à la recherche d'auteurs qui ont étudié certains aspects de la question de départ dans la perspective d'en dégager une problématique ciblée. Le souci d'en savoir plus amène à entreprendre des lectures exploratoires approfondies dont il convient de rendre compte. Ce compte rendu de l'état des connaissances actuelles sur un sujet est appelé *recension des écrits* ou *revue de la littérature*. Elle consiste à recenser les études antérieures afin de répondre à la question de départ. Comment le phénomène étudié est-il défini, décrit dans la littérature scientifique ? Quelles sont les différentes composantes, formes, variantes recensées ? Comment le phénomène a-t-il été expliqué par les chercheurs ? Y a-t-il des théories ou modèles qui proposent des explications ? Y a-t-il des recherches empiriques qui établissent des relations ? Avec quelles méthodologies ? Quelles sont les conclusions des auteurs ? Y a-t-il des éléments de convergence et de divergence ? Tels sont autant d'éléments heuristiques renseignés par la recension des écrits, laquelle joue un rôle

[2] Emile Durkheim, Le Suicide, PUF, 1993 (1ère édition 1897)

central en recherche quantitative. Elle informe quant au choix concernant l'instrumentation, les données et les techniques statistiques à utiliser dans une étude.

Dans sa recension des écrits, par exemple, Durkheim (1897) ne fait pas un inventaire exhaustif des causes du suicide, mais recense seulement les facteurs qui pèsent sur la société. Il s'agit des facteurs psychopathologiques (folie, alcoolisme), psychologiques (race et hérédité), cosmiques (climat, température), de l'imitation et des facteurs sociaux. Le célèbre sociologue interprète les statistiques de la littérature existante pour mieux évaluer l'effet de chaque facteur sur le suicide. En réalité, lorsque l'on mène une étude quantitative, la revue des écrits nécessite une certaine compréhension des statistiques et de la logique sous-jacente. Déjà à ce stade, le chercheur doit savoir procéder à la lecture critique d'articles quantitatifs. Non seulement, il doit connaître les concepts et formules statistiques couramment utilisés dans les publications. Mais il doit aussi être capable d'interpréter de façon appropriée les mesures-synthèses, graphes et tableaux qui y sont présentés, et d'en saisir les limites. Bien sûr, nous reviendrons sur ces connaissances statistiques dont la bonne compréhension facilite et rend productive la revue de la littérature quantitative.

La recension des écrits permet au chercheur de se positionner dans une discussion scientifique antérieure, et de modifier si nécessaire la question de départ. Techniquement, elle débute par des travaux portant sur le thème. S'ensuivent des travaux en lien direct avec la question de départ, puis avec le ou les cas à l'étude. On termine par un point de chute mettant en lumière ce que l'on sait et ce que l'on ignore du phénomène étudié.

1.2.3. …À la question spécifique de recherche

La recension des écrits ne consiste pas à recenser des études énoncées successivement les unes après les autres. Son intérêt principal réside plutôt en ce qu'elle fournit des éléments de réponse à la question de départ. Ces éléments de réponse doivent être bien *synthétisés*, avant de faire l'objet d'une *critique* débouchant sur une *contribution personnelle*.

On mène une recherche parce que l'état des connaissances est insatisfaisant à propos d'une question donnée. Cette insatisfaction dans les connaissances existantes permet au chercheur d'identifier des lacunes ou des avenues non exploitées et, partant, de spécifier son propre problème. Un *problème spécifique de recherche* est l'écart entre ce qu'on connait dans la littérature et ce qu'on ne connait pas encore mais qui mérite d'être connu. Après avoir recensé les facteurs explicatifs du suicide, s*achant que* ce phénomène ne peut pas être dû à des causes extra-sociales, mais plutôt à des causes proprement sociales (synthèse), *mais constatant que* les informations fournies par les statistiques sur les causes extra-sociales présumées sont lacunaires (critique), Durkheim (1897) *en déduit* que la seule méthode efficace est de chercher comment le taux de suicide varie selon le degré de cohésion sociale dans les pays européens (contribution). Voilà les termes explicités selon lesquels, le père de la sociologie spécifie son propre problème de recherche. Pour rendre explicite l'intention de contribuer à l'avancement des connaissances ou à la pratique sociale, le problème spécifique de recherche est habituellement transformé en une question. Le taux social de suicide varie-t-il en fonction du degré de cohésion sociale dans les pays européens ? Telle est la *question spécifique de recherche* qui a guidé l'étude remarquable de Durkheim (1987). Le

NE PAS COPIER

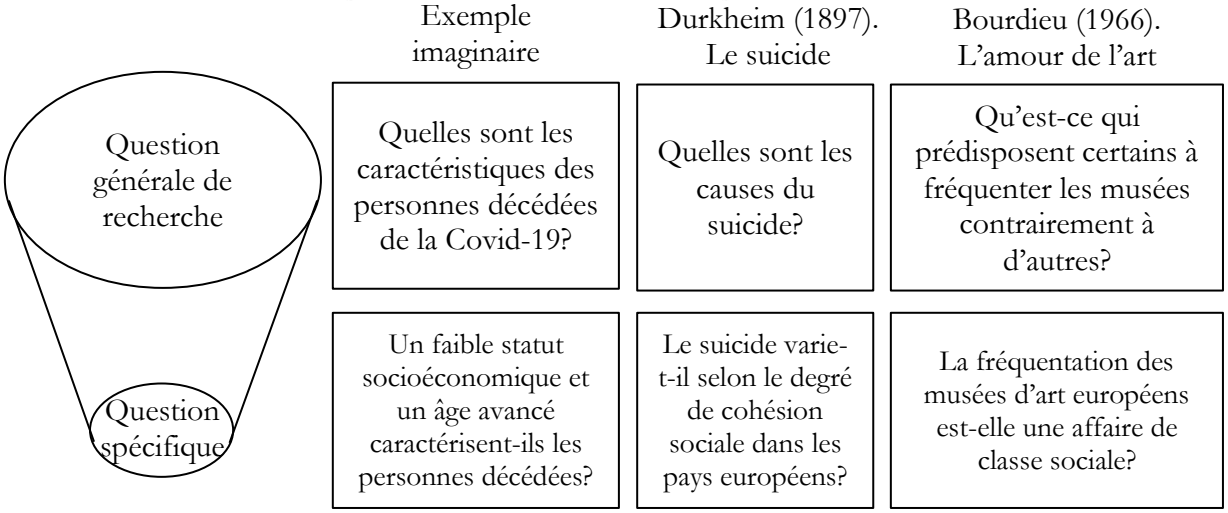
chercheur, en l'occurrence, doit pouvoir justifier l'utilité de répondre à cette *question centrale*. Dans le cas de l'étude susmentionnée, la réponse à la question spécifique permettait de mieux comprendre le rôle des concomitants sociaux dans la genèse des suicides, et ce grâce à la méthode des variations concomitantes. On peut formuler quelques sous-questions qui servent à centrer davantage le questionnement personnel du chercheur.

Sans être obsédé par la question jamais formulée, le chercheur peut s'intéresser à des faits nouveaux ou jeter un autre regard sur des faits déjà étudiés. En recherche quantitative, plusieurs situations problématiques peuvent être à l'origine d'une question spécifique pertinente, au regard des sciences sociales, aux fins de combler des lacunes.

1) Besoin de décrire un phénomène social : phénomène nouveau ou peu étudié, nouvelles variables descriptives, classificatoires ou comparatives, nouvel instrument de mesure, passer de la simple description à la classification ou comparaison, etc.
2) Besoin d'expliquer un phénomène social : nécessite de passer de la description à l'explication, théories concurrentes à tester, controverse à élucider, facteurs explicatifs nouveaux ou négligés, autre modèle d'analyse à estimer (modération, médiation…), passer de l'explication transversale à l'explication longitudinale, etc.
3) Besoin d'améliorer la méthodologie : validité des résultats à vérifier avec des données récentes ou longitudinales, nouvelles techniques statistiques à utiliser, nécessité de contrôler des relations, autre design ou stratégie de recherche, etc.
4) Besoin d'asseoir la robustesse des résultats : répliquer une étude en augmentant significativement la taille de l'échantillon, ou en changeant la population de référence, le contexte, l'approche (de qualitative à quantitative), etc.

En fin de compte, la problématique apparaît comme un processus qui consiste à passer du problème social au problème sociologique, de la question générale (départ) à la question spécifique (centrale) de recherche. Illustré à travers la figure 1.2, ce processus en entonnoir est formalisé dans un texte qui tente de justifier la question retenue et la méthode envisagée pour y répondre. Sans une question pertinente, faisable et clairement formulée, une recherche ne produira guère des savoirs utiles, même si elle repose sur des statistiques sophistiquées. La problématique doit donc être construite sur des bases solides puisqu'elle fournit le cadre théorique sur lequel s'adossent les hypothèses de recherche.

Figure 1.2. Quelques exemples illustrant la construction d'une problématique sous l'angle quantitatif



1.3. Formulation de l'hypothèse et opérationnalisation

La section précédente a mis en lumière les problèmes et questions qui structurent, d'une façon générale et spécifique, la problématique d'une recherche scientifique. À présent, le problème de fond est de savoir comment traduire cette problématique en des termes qui la rendent concrète. Tel est l'objet de cette section dont l'enjeu consiste à délimiter le cadre d'analyse d'une étude statistique. La formulation et l'opérationnalisation de l'hypothèse constituent la deuxième étape d'une recherche, notamment quantitative. L'hypothèse de recherche est décrite (1.3.1) avant de donner lieu à un premier niveau d'opérationnalisation (1.3.2), puis à un second niveau d'opérationnalisation (1.3.3).

1.3.1. L'hypothèse en recherche quantitative

L'hypothèse est un énoncé affirmatif qui répond à la question spécifique de recherche. Elle est donc logiquement et intimement liée à la problématique. Contrairement à la recherche qualitative, la recherche quantitative nécessite la formulation d'une réponse anticipée et prédictive incontournable. Dans l'étude de Durkheim (1897), l'hypothèse est faite que plus le degré de cohésion social est faible, plus le suicide augmente dans les pays européens. Une hypothèse de recherche peut être formulée sous forme d'*objectif*, notamment lorsqu'on ne prédit pas le sens de la relation étudiée. Dans l'exemple précédent, l'objectif de recherche consisterait à montrer que le suicide varie en fonction du degré de cohésion sociale.

La plupart des hypothèses formulées en recherche sociale sont de nature bivariée ou multivariée, car elles mettent en relation deux ou plusieurs phénomènes. Cependant, une hypothèse de recherche peut être univariée. Univariée, lorsqu'elle décrit un phénomène, notamment son ampleur ou évolution, ou lorsqu'elle compare une statistique observée à un paramètre théorique pour un même phénomène. Ainsi en est-il lorsqu'on affirme que le suicide a diminué au Québec depuis dix ans et qu'on tente de tester cette affirmation.

La recherche quantitative permet de tester, de façon convaincante, des hypothèses de travail. Pour ce faire, elle utilise les statistiques et adopte surtout le raisonnement déductif. La *démarche hypothético-déductive* consiste, à partir d'une ou des théories combinées, à vérifier une hypothèse par la confrontation aux faits[3]. Cette démarche confirmatoire est surtout appropriée lorsqu'il existe un cadre analytique ou théorique bien connu. Au fait, une hypothèse doit être plausible. Autrement dit, il doit y avoir des raisons suffisantes de croire que des phénomènes sont reliés entre eux. Pour ce faire, le chercheur s'appuie sur des faits et surtout des théories qui permettent d'expliquer les relations établies. Les théories, en particulier, ont un pouvoir intégrateur et structurant. Elles permettent d'intégrer des connaissances sur le monde social, tout en orientant le choix des hypothèses et les observations empiriques. Malgré le rôle indéniable des théories en recherche quantitative,

[3] En recherche qualitative, la démarche étant habituellement inductive, on cherche plutôt à induire des théories à partir de l'observation singulière des faits. Ces théories post-hoc s'apparentent à ce que le sociologue Merton (1948) appelait « théories de moyenne portée ». Dans ce contexte davantage exploratoire, la formulation apriori de l'hypothèse n'est pas une étape obligatoire pour guider le processus de recherche. Elle découle plutôt de l'analyse des données d'une expérience idiographique.

NE PAS COPIER

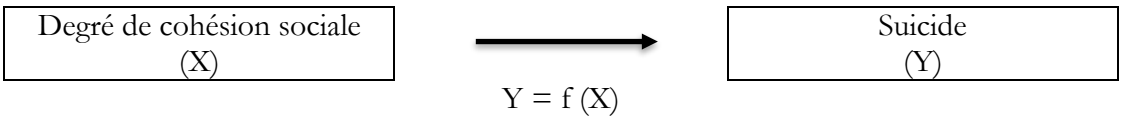
le risque est grand d'écarter tout résultat qu'elles ne peuvent ni intégrer, ni expliquer. La possibilité de découvrir des résultats insoupçonnés s'en trouve limitée en conséquence.

Non seulement, l'hypothèse de recherche doit ultimement contribuer à l'amélioration des connaissances. Mais également, elle doit pouvoir être confirmée, infirmée ou nuancée dans la réalité. Précisément, pour que l'hypothèse conduise directement à la collecte des données, elle doit être traduite en des termes qui la rendent opérationnelle dans les faits. L'opérationnalisation implique l'élaboration d'un modèle d'analyse vérifiable, c'est-à-dire un système cohérent de relations entre des concepts opératoires.

1.3.2. Premier niveau d'opérationnalisation

Le premier niveau d'opérationnalisation consiste à représenter formellement l'hypothèse de recherche, notamment en identifiant clairement la façon dont les variables sont reliées. La figure ci-dessous montre comment peut être formalisée l'hypothèse de Durkheim (1897).

Figure 1.3. Représentation formelle d'une hypothèse de recherche



La propension au suicide varie d'une société à une autre. C'est la variable centrale dont on veut expliquer justement la variation, c'est-à-dire la variable dépendante. La *variable dépendante* est donc la variable expliquée, l'effet présumé qui dépend d'une autre variable. Elle est habituellement symbolisée par la lettre Y. Quant au degré de cohésion sociale, il correspond à la variable qui explique la variation du taux de suicide. C'est la *variable indépendante*. Symbolisée souvent par la lettre X, elle est aussi appelée variable explicative, soit le facteur qui explique présumément la variable dépendante. Sur le schéma ci-dessous, la flèche symbolise la relation formelle entre les deux variables. Cette relation peut être représentée par un modèle mathématique : $Y = f(X)$. En des termes sémantiques, le suicide (Y) est fonction (f) du degré de cohésion sociale (X). Par ailleurs, Durkheim (1897) va chercher à contrôler cette relation en impliquant des *variables-contrôle* (symbolisées par Z) comme l'âge et le sexe, les personnes âgées et les hommes étant plus susceptibles de se suicider. On est dans l'univers des hypothèses multivariées, qui ne sont pas couvertes dans cet ouvrage. Nous apprendrons surtout à tester des hypothèses bivariées.

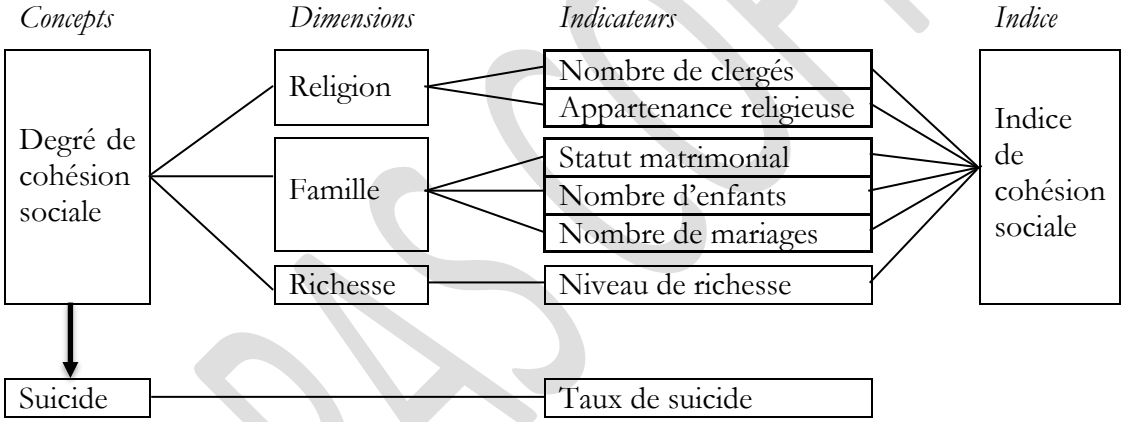
1.3.3. Deuxième niveau d'opérationnalisation

Le deuxième niveau d'opérationnalisation de l'hypothèse consiste à définir précisément les concepts, à les décomposer en dimensions, les dimensions en indicateurs, et éventuellement à fusionner les indicateurs en indices. Considéré comme l'un des ténors de la sociologie quantitativiste américaine, Lazarsfeld (1970) a établi, de façon concrète et rigoureuse, comment passer des concepts aux indicateurs empiriques.

En recherche quantitative, les concepts sous-jacents aux hypothèses doivent être précis et opératoires. Lorsque l'on conceptualise un terme, on fait un choix parmi les significations possibles. Aussi, Durkheim (1897) définit-il le suicide comme tout acte de mort résultant

directement ou indirectement de l'action d'un individu qui se sait devoir produire ce résultat. Même clairement défini, le *concept* n'est pas directement observable. Il faut donc en spécifier les *dimensions* si nécessaire, pour descendre sur l'échelle de l'abstraction, et surtout sélectionner les indicateurs pertinents. Les *indicateurs* sont les signes observables et mesurables d'un concept. Par exemple, pour mesurer le suicide dans les pays européens, Durkheim (1897) utilise le taux social de suicide comme indicateur, précisément le nombre de suicides par 100 000 habitants. C'est donc un phénomène enregistrable par les statistiques officielles. Concernant le concept de degré de cohésion sociale, l'auteur le mesure au travers de plusieurs indicateurs : nombre de clergés, appartenance religieuse, nombre d'enfants dans les familles, statut matrimonial, situation économique. Une société ayant moins de clergés ou à dominante protestante, une faible densité familiale, un faible taux de mariage, la prospérité font augmenter le taux de suicide. Durkheim (1987) aurait pu amalgamer tous ces indicateurs pour obtenir un *indice* de cohésion sociale révélateur. La figure ci-dessous résume le passage des concepts à la définition des indicateurs, puis à la construction d'un indice comme échelle de mesure synthétique.

Figure 1.4. Des concepts à l'indice en passant par les indicateurs



L'hypothèse de recherche est habituellement scindée en sous-hypothèses, lesquelles correspondent aux *objectifs spécifiques* à atteindre. Comme moyen de vérification empirique, elle suggère les dispositifs techniques à mobiliser à cet effet. Lorsqu'une hypothèse opérationnelle vise à décrire ou expliquer un phénomène quantifiable, la méthode quantitative s'avère appropriée. Encore faudrait-il bien choisir les indicateurs des concepts sous-jacents. Les statistiques, en effet, ne comprennent que le langage des indicateurs à propos desquels on collecte les données numériques.

1.4. Collecte des données numériques

La collecte des données est la troisième étape d'une recherche scientifique. En recherche quantitative, deux méthodes dominantes sont utilisées pour produire des données numériques : le devis expérimental et l'enquête par questionnaire. Ces méthodes s'inscrivent dans une stratégie visant à vérifier les hypothèses de travail (1.4.1). Les individus testés ou enquêtés sont sélectionnés selon différentes méthodes d'échantillonnage (1.4.2). Enfin, les chercheurs travaillent avec des données primaires ou secondaires (1.4.3).

NE PAS COPIER

1.4.1. Stratégie de vérification : devis expérimental vs enquête par questionnaire

Le *devis expérimental* consiste à mettre des individus dans des conditions contrôlées où l'on étudie leurs réactions en vue d'expliquer causalement un phénomène. L'expérimentation est souvent employée pour évaluer l'efficacité d'une intervention. À cet effet, le chercheur utilise un test qui lui permet de collecter des données pour vérifier si la variable indépendante (l'intervention ou non) a un effet sur la variable dépendante (ce qu'on veut tester). Supposons qu'on mette en expérience l'effet d'une nouvelle méthode d'enseignement sur la performance des étudiants en statistiques sociales. On décide de répartir aléatoirement les étudiants dans deux groupes : le groupe expérimental qui reçoît la nouvelle méthode et le groupe contrôle l'ancienne méthode. Ensuite, on teste la performance pour chacun des groupes et on les compare. Si la nouvelle méthode est efficace, le groupe expérimental aura de meilleures notes. Ici, la variable indépendante est *manipulée* ou provoquée, puisqu'on peut assigner la nouvelle méthode à tel ou tel autre groupe. On est donc dans l'univers expérimental. En assignant aléatoirement les étudiants au groupe expérimental ou au groupe contrôle (randomisation), l'expérimentateur obtient des groupes semblables et contrôle ainsi naturellement les facteurs susceptibles d'influer sur la relation de cause à effet. Lorsqu'il n'est pas possible de répartir les sujets de façon aléatoire et qu'on se contente de tester des groupes préexistants mais dont les caractéristiques sont similaires (matching), on parle de *quasi-expérimentation*. Quoiqu'il en soit, les conditions pour mener une étude expérimentale ou quasi-expérimentale sont rarement réunies en sciences sociales où le chercheur n'est pas toujours en mesure d'intervenir dans le cours naturel des choses. La plupart des variables d'intérêt ne peuvent pas faire l'objet d'une manipulation, en raison de la nature des faits sociaux et pour des raisons éthiques. On pourrait cependant procéder à une étude observationnelle.

L'enquête par questionnaire. Elle consiste à interroger directement les individus à propos de leurs comportements, attitudes, opinions, dans la perspective de décrire et d'expliquer un phénomène social. Lorsqu'un chercheur considère le sexe pour expliquer les différences dans la performance des étudiants en statistiques sociales, il ne peut manipuler ce facteur. Il ne peut aucunement interchanger les hommes et les femmes afin d'observer leurs réactions. Tout ce que le chercheur peut faire, c'est de désigner les hommes comme étant des hommes, les femmes comme des femmes, et de leur poser des questions concernant les notes obtenues. Ici, la variable indépendante, le sexe, est simplement *désignée* ou invoquée. Dans ce contexte, on mène une étude observationnelle, une enquête par questionnaire précisément. Le chercheur ne fait qu'observer in vivo le phénomène étudié, sans être en mesure d'intervenir pour modifier les caractéristiques des participants.

Le questionnaire est l'outil de mesure le plus fréquemment utilisé en sciences sociales. Même s'il repose souvent sur des questions fermées (de type oui/non, choix multiple) ou semi-fermées (choix multiple incluant autre), il peut inclure des questions ouvertes (non structurées). Le formulaire de questions systématiques peut être autoadministré lorsqu'il est complété par les répondants eux-mêmes sur Internet ou par correspondance. Il peut être aussi administré par téléphone ou en face à face notamment avec la méthode CAPI (Computer Assisted Personnal Interviews) qui permet à l'enquêteur de mener des

interviews assistées par l'usage de tablettes. Dans ce dernier cas précisément, la présence de l'interviewer peut avoir une influence sur les propos des répondants.

Je voudrais maintenant discuter de deux points essentiels. D'une part, les répondants ont tendance à montrer une image avantageuse d'eux-mêmes. C'est ce que les sociologues appellent *désirabilité sociale*. De sorte que lorsqu'on pose des questions dont les réponses relèvent du socialement acceptables ou non, il est important de commencer par une introduction qui laisse entrevoir que toutes les réponses sont acceptables. Par exemple, pour mesurer le comportement des citoyens vis-à-vis de la pandémie, il n'est pas approprié de demander directement : *Au cours de la dernière semaine, avez-vous oui ou non respecté les mesures barrières édictées par le gouvernement pour contrer la Covid-19 ?* On gagnerait plutôt à introduire en ces termes : *Certaines personnes respectent les consignes. D'autres non pour diverses raisons compréhensibles. Et vous*… ? D'autre part, beaucoup de phénomènes sociaux ne sont pas directement observables et mesurables auprès des répondants. Cependant, les chercheurs utilisent des *variables proxy*, comme le revenu pour mesurer le statut socioéconomique, ou des *échelles de mesure*. L'usage des échelles est très courant en psychologie, en sociologie et en science politique. Il se justifie par la difficulté à mesurer un concept par une seule question. Les répondants ne sont pas en effet censés se substituer au travail du chercheur. À titre d'illustration, pour mesurer empiriquement le postmatérialisme dans une société, on ne posera pas une question générale du genre : *Selon vous, êtes-vous matérialiste ou postmatérialiste ?* Car ce concept éminemment sociologique est abstrait et peut-être compris différemment par les répondants eux-mêmes. Le politologue Inglehard (1970) mesure plutôt le postmatérialisme, qu'il a théorisé, par cette question à choix multiple : *Si vous deviez choisir parmi les choses suivantes, lesquelles sont les deux qui vous semblent les plus désirables ?*

1.- Maintenir l'ordre dans la nation
2.- Donner plus de parole aux gens dans les décisions politiques importantes
3.- Lutter contre la hausse des prix
4.- Protéger la liberté d'expression.

Si les répondants adhèrent fortement aux valeurs matérialistes (sécurité, bien-être matériel, etc.), ils auront tendance à cocher les choix 1 et 3. S'ils sont plutôt postmatérialistes (reconnaissance, affirmation, etc.), ils choisiront 2 et 4 comme options. Par le biais de ces indicateurs, le chercheur pourra mesurer le postmatérialisme en tant que système de valeurs relatif au désir de satisfaction des besoins non matériels et non consuméristes.

Par ailleurs, il est primordial de savoir qu'il y a un raisonnement statistique qui sous-tend l'utilisation de l'enquête par questionnaire (et du devis expérimental par extension). Au moins trois principes sous-jacents sont à retenir :

1) *Principe épistémologique* : Pour étudier un phénomène social donné, le chercheur peut interroger l'individu puisque celui-ci fait partie intégrante d'un tout, la société, qui influe sur ses comportements, ses attitudes et ses opinions ;
2) *Principe théorique* : Les tendances majoritaires qui se dégagent des réponses d'une fraction d'individus apportent des informations probantes pouvant aider à décrire et expliquer des phénomènes sociaux ;

NE PAS COPIER

3) *Principe statistique* : Les informations issues d'une fraction d'individus choisis de façon appropriée peuvent être généralisées à la population de référence.

1.4.2. L'échantillonnage

Dans le cadre d'une expérimentation ou d'une enquête, il n'est pas toujours possible de tester ou d'interroger tous les individus de la population d'intérêt. En revanche, il est possible de prélever un échantillon à partir de cette population. Cette procédure est appelée échantillonnage. Habituellement, on distingue les méthodes d'échantillonnage aléatoire et les méthodes d'échantillonnage non aléatoire. Leur distinction comporte d'importantes implications en statistiques inférentielles. C'est pourquoi nous passerons en revue et comparerons ces méthodes d'échantillonnage dans le chapitre 7. À ce stade, il nous semble important d'en avoir un simple aperçu.

L'*échantillonnage aléatoire*, ou probabiliste, consiste à choisir par tirage au sort les individus à enquêter ou à soumettre à un test. On est ici dans l'univers du hasard, au sens statistique du terme. Cette procédure nécessite le respect d'au moins deux règles. La première règle consiste à disposer d'une base de sondage, soit la liste numérotée et exhaustive des individus constituant la population étudiée. La deuxième règle veut que le tirage au sort des individus se fasse à partir d'un générateur de nombres aléatoires (table, logiciel échantillonneur). Les individus peuvent être choisis au hasard à l'intérieur de groupes homogènes (strates, grappes). Bien qu'il produise à la longue un échantillon représentatif de la population, l'échantillonnage aléatoire nécessite une connaissance fine de la population parente.

L'*échantillonnage non aléatoire*, ou empirique, ne nécessite pas que la population soit nommément connue. Il consiste à choisir les individus de façon délibérée selon des critères fixés par le chercheur. Ce dernier choisit ainsi ceux qui satisfont à ces critères tantôt arbitraires tantôt raisonnables. L'enjeu est d'obtenir un échantillon caractéristique de la population en l'absence d'une base de sondage. Cette méthode d'échantillonnage ne fait pas intervenir le hasard. On l'appelle aussi méthode de convenance, puisque le chercheur peut choisir des volontaires, des cas-typiques ou des personnes rencontrées accidentellement. L'échantillonnage non aléatoire est certes facile à mettre en place. Il n'en demeure pas moins qu'il tend à produire un échantillon sur lequel il est difficile d'appliquer les principes de l'inférence statistique, comme nous le verrons.

1.4.3. Données primaires vs Données secondaires

Quand on recueille de nouvelles données dans le cadre d'un processus de recherche, on obtient des données primaires. Les *données primaires* sont les données recueillies par le chercheur lui-même suivant son propre plan de recherche. Elles sont alors produites au moyen d'une enquête par questionnaire ou d'un devis expérimental. Certes, le chercheur maîtrise davantage les conditions de production des données primaires. Mais, la production présente d'importants coûts en termes financiers, humains et matériels. Aussi lorsqu'on doit vérifier une hypothèse de travail, le recours aux données secondaires apparait comme une alternative efficiente. Les *données secondaires* sont les données qui préexistent au chercheur et dont il se sert pour obtenir des statistiques. Elles présentent des avantages indéniables : 1) elles sont déjà produites ; 2) elles sont rapidement accessibles et

relativement peu dispendieuses. Toutefois, lorsqu'elles existent et qu'elles sont actualisées, les données secondaires ne font pas toujours l'affaire des utilisateurs. Elles peuvent ne pas inclure des variables d'intérêt ou des méta-informations sur la méthodologie employée.

Très souvent négligées, les données secondaires abondent pourtant. Pour expliquer le suicide dans les pays européens par des régularités sociologiques, Durkheim (1897) s'est principalement servi de publications statistiques officielles datant de la fin du 19e siècle. De nos jours, les sources de données secondaires se sont diversifiées. Par exemple, depuis le début de la crise sanitaire, les médias présentent des données en lien avec le nombre d'infectés, d'hospitalisations, de décès, de personnes guéries, etc. Une fois publiées, ces données devenaient des matériaux secondaires précieux pour les chercheurs. Se pose alors la question de leur accès. Grosso modo, on distingue les producteurs et les distributeurs. Les *producteurs*, comme Statistique Canada, sont les organismes qui produisent des données statistiques. Pour informer les politiques publiques, Statistique Canada collecte d'abord des données primaires, qui deviennent dès leur publication des données secondaires pour les chercheurs. Il est souvent difficile d'accéder directement à ces données. En revanche, l'accès peut se faire facilement à travers les *distributeurs*, soit les organismes ou plateformes qui gèrent des banques de données et les distribuent gratuitement aux utilisateurs. On note partout dans le monde l'émergence d'initiatives de démocratisation de l'accès aux données statistiques (open data). A titre d'information, voici une énumération sommaire de quelques données secondaires utilisées en sciences sociales.

Au Canada, la plateforme ouverte des données publiques[4] met à la disposition des chercheurs de multiples bases de données, dont celle de l'enquête sociale générale (ESG). L'ESG est une importante enquête sociologique pilotée par Statistique Canada. Elle porte sur les tendances sociales, les conditions de vie et le bien-être des Canadiens. De même, les systèmes interuniversitaires de gestion de données, comme Odési, permettent d'offrir aux universitaires les statistiques comme ressources électroniques. Ces systèmes gèrent de vastes données d'enquêtes canadiennes. *En France*, le réseau Quételet[5] offre un portail qui permet de chercher et d'accéder aux données françaises en sciences sociales. Ces données sont issues des statistiques publiques et de toutes sortes d'enquêtes menées en France.

Au niveau continental, les enquêtes baromètres diffusent des données libres issues de sondages d'opinion publique sur des questions économiques, politiques et sociaux pertinentes pour les décideurs (afrobaromètre, Latino-baromètre, baromètre asiatique, baromètre arabe, baromètre eurasiatique[6], et eurobaromètre[7]). *À l'international*, les banques de données de la Banque mondiale[8] permettent un accès gratuit et libre à des statistiques sur le développement dans le monde. Il en est ainsi des bases de données de l'Organisation des Nations Unies (ONU)[9]. Le *World Values Survey* (WVS)[10] est un programme international de

[4] https://ouvert.canada.ca/data/fr/dataset?q=, consulté le 21 janvier 2021.
[5] http://bdq.quetelet.progedo.fr/fr/Accueil, consulté le 11 janvier 2021.
[6] https://www.globalbarometer.net/, consulté le 11 janvier 2021.
[7] https://www.gesis.org/en/eurobarometer-data-service/search-data-access/data-access, consulté le 11 janvier 2021.
[8] https://donnees.banquemondiale.org/, consulté le 11 janvier 2021.
[9] https://www.un.org/en/databases/index.html#stats, consulté le 11 janvier 2021.
[10] http://www.worldvaluessurvey.org/wvs.jsp, consulté le 11 janvier 2021.

NE PAS COPIER

recherche dont les données sont largement accessibles. Ce programme scientifique est consacré à l'étude des valeurs et leur impact sur le développement des sociétés.

Somme toute, les données secondaires présentent des avantages et limites qu'il convient de connaître afin de déterminer leur pertinence eu égard à une étude statistique. Les données primaires, quant à elles, ne sont essentielles aux chercheurs que lorsque des données secondaires probantes et actualisées n'existent pas. Qu'elles soient de première main ou de seconde main, individuelles ou agrégées, les données collectées doivent faire l'objet d'une analyse statistique pour être intelligibles, communicables et utiles.

1.5. Analyse des données numériques

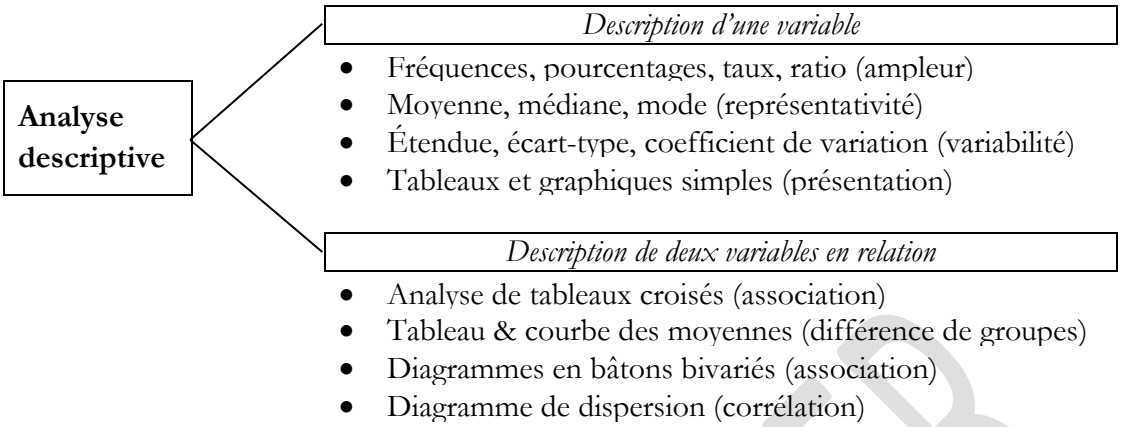
Au terme de la collecte des données, le chercheur s'est retrouvé devant un amas de matériaux bruts, divers et disparates, qu'il convient de traiter afin de les rendre exploitables. Le traitement consiste à transformer les données en des termes qui les rendent intelligibles au raisonnement statistique. Cette opération prépare l'analyse proprement dite qui correspond à la quatrième étape d'un processus de recherche. L'analyse consiste à générer des informations qui aident à vérifier des hypothèses de travail ou supporter une décision. Nous verrons plus loin comment traiter statistiquement des données numériques : codage, recodage, validation, etc. À présent, il nous semble important que le lecteur ait une vue d'ensemble des méthodes d'analyse statistique (1.5.1) à aborder dans ce manuel, tout en étant en mesure de distinguer l'analyse principale de l'analyse secondaire (1.5.2) et de savoir comment les résultats statistiques sont habituellement présentés (1.5.3).

1.5.1. Les méthodes d'analyse statistique

Une fois les opérations de saisie et de traitement complétées, l'analyse comme telle peut s'amorcer. Deux grandes familles de méthodes statistiques, introduites dans cette section, sont disponibles à cet effet : l'analyse descriptive et l'analyse inférentielle.

L'*analyse descriptive* est l'étape préliminaire à toute analyse statistique. Elle consiste à résumer, décrire les données afin de les rendre informatives. L'analyste peut décrire une population ou, de façon indépendante, un échantillon. Mais comme les données de population sont rarement disponibles, l'analyse descriptive porte plus souvent qu'autrement sur des données d'échantillon. Elle peut être univariée, bivariée, ou multivariée. *Univariée*, lorsqu'elle cherche à décrire les données des variables prises une à une. *Bivariée*, lorsqu'elle décrit les données de deux variables reliées entre elles. *Multivariée*, lorsqu'elle décrit les relations entre trois variables ou plus. Seules les analyses univariées et bivariées occuperont notre attention dans cet ouvrage. Elles sont illustrées dans la figure ci-dessous.

Figure 1.5. Aperçu des techniques d'analyse statistique descriptive



Si sophistiquée soit une étude quantitative, elle débute presque toujours par une *analyse descriptive univariée* qui produit une photographie des données de chaque variable. Illustrons maintenant l'utilité de ce genre d'analyse en partant d'un problème de recherche imaginaire. Imaginons qu'un chercheur étudie les inégalités salariales au Canada et qu'il veuille connaître la distribution du revenu de 1000 adultes sélectionnés. On obtient des données individuelles relatives à 1000 revenus qui n'offrent aucune cohérence interne. Il faut donc les traduire en informations concises facilement utilisables. Pour y arriver, les données doivent être résumées à l'aide de *statistiques descriptives* appropriées. Concrètement, le chercheur peut décrire le revenu en utilisant la moyenne, un graphique ou un tableau présentant le pourcentage de revenus faible, moyen et élevé. C'est l'objet de l'analyse descriptive qui obéit au principe de réduction des données. L'enjeu est d'obtenir des résumés quantitatifs qui permettent de décrire le profil des participants à une étude.

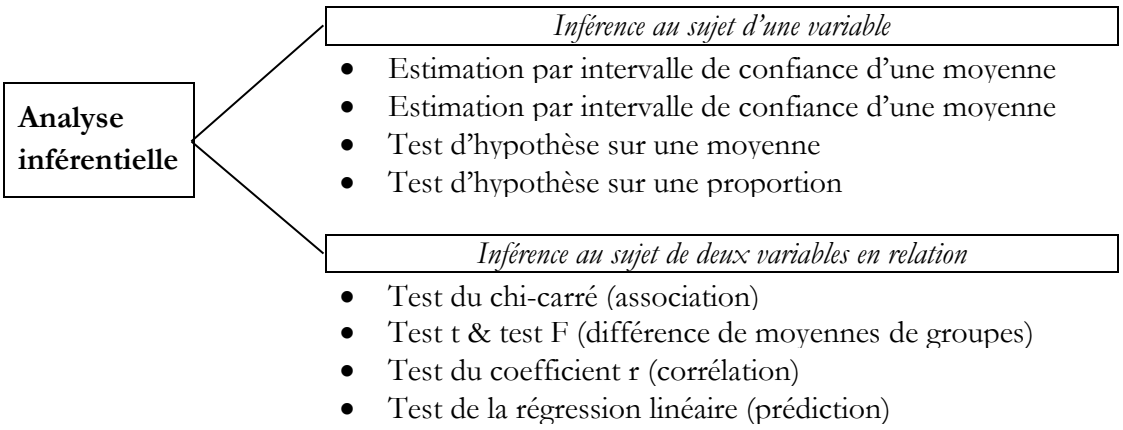
L'analyse descriptive univariée aide à valider les données. Au fait, une fois les variables décrites individuellement, on cherche ensuite à détecter des associations deux à deux entre elles. Cette *analyse descriptive bivariée* permet de décrire les relations entre deux variables, en comparant plusieurs profils d'individus. Pour expliquer les inégalités salariales au Canada, supposons qu'on mette en relation le sexe et le revenu des 1000 adultes enquêtés. Une façon de résumer les données consiste à utiliser un tableau des moyennes qui compare le revenu moyen des femmes et des hommes, ou un tableau croisé qui présente les pourcentages de femmes et d'hommes quant aux niveaux de revenu faible, moyen et élevé.

Qu'elle soit univariée ou bivariée, l'analyse descriptive est très utile à des fins exploratoires. Mieux, toute une recherche sociale peut être menée à des fins uniquement descriptives. Il en est ainsi lorsqu'on explore un phénomène nouveau ou peu étudié comme la pandémie de la Covid-19. Ou lorsqu'il s'agit de décrire les caractéristiques d'un problème social et les facteurs qui y sont associés. Ou encore lorsqu'on travaille avec un échantillon de petite taille qui limite toute possibilité de généralisation. Toutefois, les études uniquement descriptives sont plutôt rarement rencontrées en sciences sociales. Dans bien des cas, après avoir réduit les données univariées et bivariées en informations claires et fines, on cherche à généraliser ces informations à la population dont provient l'échantillon. C'est l'objet de l'*analyse inférentielle*. Elle consiste à inférer à la population de référence les résultats obtenus. L'inférence statistique obéit donc au principe de généralisation des informations, et cherche

NE PAS COPIER

à évaluer l'incertitude associée à cette généralisation. Par exemple, il faut s'assurer que le revenu moyen des 1000 adultes canadiens sondés est statistiquement significatif au niveau de toute la population canadienne étudiée en utilisant l'estimation par intervalle de confiance[11]. Il s'agit là d'une *analyse inférentielle univariée*. Mais en recherche sociale, on utilise les *statistiques inférentielles* surtout pour tester des hypothèses relatives à des relations bivariées, comme celle entre le sexe et le revenu. Pour savoir si la différence entre les hommes et femmes eu égard à leur revenu moyen est significative ou non, on utilise le test t. Alors que le test du chi-carré est appliqué pour valider s'il y a des différences significatives entre les pourcentages des femmes et hommes quant aux niveaux de revenu faible, moyen et élevé. La figure ci-dessous visualise les différentes statistiques inférentielles à aborder.

Figure 1.6. Aperçu des techniques d'analyse statistique inférentielle



Cette section fait un simple survol des différentes techniques d'analyse statistique univariée et bivariée que nous ne développons pas ici. Tout l'ouvrage, en effet, tourne autour de ces techniques ; la deuxième partie de l'ouvrage étant consacrée à l'analyse descriptive, et la troisième partie à l'analyse inférentielle. En attendant, il est important de distinguer l'analyse selon qu'elle est principale ou secondaire et d'en tirer quelques implications.

1.5.2. Analyse principale vs Analyse secondaire

L'*analyse principale* porte sur les données se rapportant aux hypothèses de recherche. Elle est effectuée en fonction des variables du modèle d'analyse. Son objectif est de fournir des éléments d'information qui aident à confirmer, infirmer ou nuancer les hypothèses. L'*analyse secondaire*, quant à elle, s'adresse aux données s'inscrivant en dehors du modèle d'analyse. Elle s'applique aux variables n'ayant pas donné lieu à une ou des hypothèses, mais qui demeurent importantes pour éclairer le problème étudié. Par exemple, si l'analyse principale révèle une relation significative entre le sexe et le revenu, on peut ensuite se demander si le sexe a un effet sur la scolarité, un indicateur de statut socioéconomique comme le revenu. Ou bien si la relation entre le sexe et le revenu persiste avec la même intensité chez les personnes fortement scolarisées et moins scolarisées ou chez les personnes âgées et moins

[11] Vous êtes, peut-être, familiers avec les sondages électoraux qui fournissent l'exemple typique d'utilisation des statistiques inférentielles. Lorsqu'un sondage d'opinion publique estime à 51% la proportion de citoyens ayant l'intention de voter pour un candidat, il donne souvent la marge d'erreur associée à cette estimation. La marge d'erreur permet de calculer l'intervalle de confiance à un certain niveau (95% par exemple), comme nous le verrons.

âgées. L'analyse secondaire permet ainsi d'apprécier la robustesse des relations bivariées analysées, et d'ouvrir des perspectives sur des analyses multivariées plus poussées. Elle est souvent utilisée pour comparer les résultats suivant les caractéristiques sociologiques des participants à l'étude. Cela dit, seuls les résultats secondaires suffisamment significatifs sont retenus. Au fond, l'analyse secondaire complète, enrichit et nuance l'analyse principale.

La plupart des analyses statistiques visent à produire uniquement des informations qui cherchent à répondre précisément à la question spécifique d'une recherche. C'est sans doute l'une des principales critiques adressées à l'encontre de la recherche quantitative. Pourtant, les résultats qui ne se rapportent pas directement à la vérification des hypothèses peuvent fournir des éléments de réponses éclairants. L'analyse secondaire donne l'occasion de découvrir des résultats inattendus, un peu comme en recherche qualitative. Elle peut occasionner la *sérendipité*, un concept qui occupe une place centrale dans la pensée du sociologue Merton (1948). Dans son entendement, la *sérendipité* se rapporte à la découverte de faits insoupçonnés, pouvant amener à émettre une nouvelle hypothèse, à réviser ou nuancer une hypothèse existante. En réalité, seule l'analyse secondaire est à même de produire des connaissances de cette nature, c'est-à-dire de découvrir de nouvelles caractéristiques ou relations. Toute recherche quantitative, bien qu'elle repose essentiellement sur l'analyse principale, devrait réserver une place à l'analyse secondaire. Elle devrait aussi accorder un soin particulier à la façon dont les résultats sont présentés.

1.5.3. Présentation des résultats statistiques

L'analyse statistique des données produit des résultats souvent présentés sous forme de textes, de tableaux ou de figures. Les techniques tabulaires et graphiques aident à communiquer des informations de la manière la plus claire possible, facilitant ainsi leur interprétation. Elles peuvent présenter la même information statistique, mais sur une base différente. Construits sous forme de lignes et de colonnes, les tableaux contiennent des statistiques précises donnant ainsi un aperçu complet des données. Habituellement, leur usage est généralisé dans le milieu de la recherche. Souvent animés par un souci de précision et de rigueur, les chercheurs préfèrent chercher dans le détail des chiffres.

Cependant, les figures (graphes, schémas, cartes, etc.) sont surtout utilisées dans le milieu professionnel en raison de leurs propriétés visuelles. Elles contiennent des éléments visuels qui permettent d'avoir une vue d'ensemble rapide d'un phénomène et les variations autour. Pensons aux graphiques tape à l'œil présentés dans les médias ou en administration. Les décideurs sont constamment mus par le souci de prendre rapidement des décisions sur la base de la lecture d'informations graphiques. Tout comme les médias s'efforcent à amener le public à se faire une opinion rapide sur les grands enjeux qui traversent la société. Les descriptions graphiques sont très utiles pour observer facilement l'ampleur, la représentativité et la variabilité d'un phénomène, et pour comparer des distributions.

S'il n'est pas en général approprié de présenter une même information à la fois sous forme tabulaire et graphique, il est parfois intéressant de varier l'usage de ces techniques de présentation selon la nature de l'information à communiquer. Lorsqu'un chercheur veut illustrer l'évolution à travers le temps du taux de suicide, peut-être gagnerait-il à utiliser un

NE PAS COPIER

graphique. En revanche, l'usage d'un tableau serait probablement mieux indiqué pour fournir les pourcentages exacts des groupes de deux distributions. Quoi qu'il en soit, on présentera d'abord les résultats de l'analyse principale avant ceux de l'analyse secondaire. Lorsque les résultats sont présentés d'une façon organisée, ils deviennent plus lisibles. Le lecteur doit savoir prendre du recul sur ces techniques de présentation de l'information statistique, laquelle peut faire l'objet d'une instrumentalisation quelconque.

Somme toute, le chercheur ne réussira à donner un sens aux données que s'il a une bonne compréhension des principes de l'analyse statistique. Ces données, elles-mêmes, ne servent pas à grand-chose si elles ne sont pas exploitées à l'aide de techniques statistiques appropriées. De nos jours, avec le développement de l'informatique, les ordinateurs offrent divers outils logiciels puissants pour analyser automatiquement de grands ensembles de données numériques. De telle sorte que les analystes bénéficient plus de temps à consacrer à la présentation et, surtout, l'interprétation des résultats qui en découlent.

1.6. Interprétation des résultats statistiques

Certes les statistiques représentent des outils incontournables en recherche quantitative, mais elles ne constituent pas une fin en soi. Au fait, comme elles ne parlent pas d'elles-mêmes, il convient de les faire parler en les interprétant. L'interprétation des résultats statistiques est l'étape ultime de la recherche quantitative. Elle consiste à assigner une signification cohérente aux résultats. L'interprétation statistique est à distinguer de l'interprétation théorique/sociologique (1.6.1). Tout l'ouvrage tourne autour de l'interprétation statistique. Quant à l'interprétation théorique/sociologique, elle se décline au moins en deux formes : la discussion (1.6.2) et la critique des résultats (1.6.3).

1.6.1. Interprétation statistique vs interprétation théorique/sociologique

Quand on se contente de donner aux résultats une signification cohérente au regard des techniques statistiques utilisées, on les interprète statistiquement. L'*interprétation statistique* consiste à révéler ce que disent les chiffres et ce qu'ils suggèrent. Son succès dépend de la capacité du chercheur à comprendre le raisonnement statistique. Voici une façon d'interpréter statistiquement les résultats de l'étude sur les inégalités salariales de genre :

En 2006, les hommes avaient un revenu (50000$) supérieur à celui des femmes (34000$). Le test t de Student suggère que la différence est statistiquement significative dans la population canadienne (t= 30,09 ; p<0,05). Toutefois, l'effet du sexe sur le revenu est de faible taille (E^2=.06).

Une fois les résultats interprétés statistiquement, il faut les interpréter théoriquement, voire sociologiquement. L'*interprétation théorique* consiste à conférer un sens théorique ou sociologique aux résultats, en cherchant à les expliquer et à en dégager des implications par rapport au contexte étudié, aux autres contextes, à la pratique sociale, et aux recherches futures. Si dans l'interprétation statistique, le chercheur s'en tient aux faits chiffrés, dans l'interprétation théorique, il fait appel à son sens de jugement, à sa vision de la société. Cependant, le chercheur doit dégager le plus possible le sens objectif des résultats obtenus. Ses pairs devraient raisonnablement accepter le bien-fondé des explications et implications avancées. Comme tout l'ouvrage est consacré à l'interprétation statistique, il nous semble pertinent de nous épancher sur l'interprétation théorique/sociologique.

1.6.2. Discussion des résultats

La discussion des résultats est l'élément fondamental de l'interprétation théorique/sociologique. Elle consiste à comparer les résultats obtenus aux résultats attendus, et à expliquer les conclusions qui en résultent. Si les résultats obtenus correspondent aux résultats attendus, on parle de découvertes positives. Les *découvertes positives* renvoient aux conclusions d'une étude qui confirme les hypothèses de travail, qui ainsi deviennent des hypothèses théoriques. Encore faudrait-il expliquer ces conclusions par un retour sur la problématique. Pour ce faire, le chercheur doit faire preuve d'imagination sociologique en s'interrogeant sur les motivations, processus et contextes qui conduisent à l'effet observé ou qui interviennent dans l'apparition du phénomène étudié. Il peut se référer à des théories, des faits, ou au bon sens. Souvent présents dans la recension des écrits, ces éléments explicatifs doivent être articulés ici aux résultats. Par exemple, nous constatons que l'hypothèse d'une disparité salariale selon le genre est confirmée. Cela peut s'expliquer, entre autres, par la théorie de la socialisation ou la discrimination. Les femmes sont socialisées de telle sorte qu'elles investissent des emplois peu rémunérateurs. Nous savons également qu'elles sont victimes de discrimination sur le marché du travail. De même, Durkheim (1897) explique l'effet du degré d'intégration des groupes sociaux sur le suicide par l'importance des structures sociales en tant qu'elles déterminent l'action humaine. Lorsqu'une hypothèse est vérifiée et expliquée, elle devient un élément possible de théorie. Et donc, non seulement les résultats doivent être expliqués par une théorie, mais aussi une théorie si pertinente soit-elle, doit être testée par une recherche.

Lorsque les résultats obtenus s'écartent plutôt des résultats attendus, on est en présence de découvertes négatives. Les *découvertes négatives* sont les conclusions d'une étude qui infirme les hypothèses de travail. Dans ce cas, le chercheur est censé expliquer l'écart constaté. Par exemple, si l'hypothèse des disparités salariales selon le genre était infirmée, on pourrait l'expliquer par le contexte de l'étude (politiques de discrimination positive) ou des problèmes méthodologiques (échantillon non représentatif). Le chercheur n'a pas toujours à être déçu des découvertes négatives : il aura apporté des correctifs et préparé le terrain pour de nouvelles recherches. Si des données fiables contredisent, en tout ou en partie, les théories sur lesquelles le chercheur s'est appuyé, cela signifie qu'il faut réviser ou nuancer ces théories. De même, certaines idées que nous jugions plausibles peuvent ne pas se vérifier dans la réalité, car reposant simplement sur des stéréotypes ou préjugés. Par exemple, les immigrants sont souvent perçus comme des gens moins instruits provenant de pays où le niveau d'instruction demeure faible. Pourtant, plusieurs statistiques ont démontré qu'ils sont plus instruits que les Canadiens de naissance. Cela peut s'expliquer notamment par une politique canadienne d'immigration sélective.

1.6.3. Critique des résultats

Une fois l'hypothèse confirmée et sa plausibilité expliquée, il faut s'assurer qu'aucune lacune majeure n'est venue entachée la validité des résultats. La *validité* permet d'apprécier le degré de confiance que l'on peut avoir à propos des conclusions d'une étude. Le chercheur

NE PAS COPIER

procède alors à la critique des résultats par un retour réflexif sur la méthodologie. Ce retour porte au moins sur deux critères : la validité interne et la validité externe.

La *validité interne* permet d'évaluer la certitude ou crédibilité des résultats tirés de l'analyse des données. Dans le cas de l'hypothèse des inégalités salariales de genre, peut-on affirmer sans risque de se tromper que le sexe influe sur le revenu ? (conclusion interne). D'un côté, peut-on avoir confiance aux données collectées et aux outils de collecte utilisés ? Dans les enquêtes par questionnaire, les participants ont tendance à déformer leurs réponses en fonction des attentes du chercheur ou de la société. De telle sorte qu'il est possible d'observer un *effet de désirabilité sociale*. Par exemple, il est connu que les hommes sont plus susceptibles de gonfler leur revenu que les femmes. Ce biais potentiel peut contribuer à augmenter artificiellement les écarts de revenus. Un effet comparable est observé dans les devis expérimentaux : l'*effet placebo*. Lorsque les participants savent qu'ils sont testés, par exemple en recevant un médicament, faux ou pas, ils peuvent orienter leurs réactions dans le sens de l'efficacité attendue. Dans ce cas, il peut être difficile d'affirmer avec certitude si l'effet du traitement est dû ou non à l'intervention. Tous ces effets (désirabilité sociale, placebo) peuvent occasionner des *erreurs biaisées de mesure*, tout comme le dispositif instrumental lui-même. Lorsqu'un questionnaire comporte des questions imprécises ou orientées, il produit nécessairement des réponses peu crédibles. C'est pourquoi les questions doivent adresser des faits précis et être formulées de façon à suggérer aux répondants que toutes les réponses sont possibles. Dans le cas de l'expérimentation, les conditions du test doivent être les mêmes pour tous les participants.

D'un autre côté, y a-t-il des variables additionnelles qui ont pu confondre la ou les relations étudiées ? Par exemple, à scolarités égales, les hommes ont-ils un revenu supérieur aux femmes ? Il est plausible qu'une scolarité plus élevée des hommes contribue à creuser davantage les écarts de revenus avec les femmes. La scolarité interprète potentiellement la relation entre le sexe et le revenu. Lorsqu'un chercheur rapporte un coefficient de corrélation entre le taux d'analphabétisation et la prévalence du VIH-Sida dans les pays, il faut savoir qu'il s'agit là d'une relation d'association et non de causalité. La région économique explique potentiellement cette relation fallacieuse : les pays pauvres sont plus susceptibles de connaître à la fois un taux d'analphabétisme et une prévalence du VIH-Sida élevés. Le chercheur doit tenir compte des variables intermédiaires ou antécédentes pour mieux évaluer la qualité de son plan d'analyse. En sciences sociales, en effet, les phénomènes sont imbriqués les uns aux autres, les causes étant multiples et en chaîne. Cet enchevêtrement rend difficile toute entreprise consistant à démontrer la plausibilité d'une relation bivariée. Vous apprendrez dans un cours avancé à procéder à l'élaboration, c'est-à-dire à inclure une ou des variables additionnelles dans une telle relation.

La *validité externe,* un autre critère de scientificité, permet d'évaluer le caractère généralisable des résultats d'un échantillon à la population dont provient cet échantillon. Précisément, peut-on généraliser de façon convaincante la relation entre le sexe et le revenu à toute la population canadienne étudiée ? (conclusion externe). Lorsqu'on rapporte une différence entre les revenus moyens des hommes et des femmes au niveau d'un échantillon, le lecteur doit savoir qu'il s'agit là de statistiques descriptives et que celles-ci n'informent pas sur les vrais paramètres. Le test *t* permet de valider l'hypothèse d'une disparité salariale selon le

genre. Encore faudrait-il que l'échantillon choisi soit représentatif de la population. Nous verrons que la représentativité de l'échantillon dépend au moins de sa taille et de son caractère aléatoire. On parle d'*erreur biaisée d'échantillonnage* lorsqu'une inférence repose sur un échantillon non aléatoire. Il faut toutefois savoir que même si un échantillon est aléatoire, il comporte toujours une erreur associée au fait de sélectionner tels individus plutôt que d'autres. C'est cette *erreur aléatoire d'échantillonnage* qu'on tente d'estimer en inférence statistique. Par exemple, un paramètre de type $p<0.05$ indique qu'on peut avoir confiance à au moins 95% en la généralisation d'une relation, la probabilité de risque d'erreur aléatoire étant seulement de 5%. Nous reviendrons largement sur ces concepts fondamentaux associés aux statistiques inférentielles. Par ailleurs, la validité externe des résultats est nécessairement affectée par les conditions de collecte des données. En ce sens, les données statistiques officielles utilisées par Durkheim (1897) dans son analyse du suicide n'étaient pas forcément homogènes d'un pays à l'autre.

En fin de compte, la validité externe est à distinguer de la *portée* des résultats. La validité externe concerne la généralisation des résultats obtenus d'un échantillon à la *population étudiée*. Elle pose une question qui peut être résolue sur le plan statistique : Dans quelle mesure une relation, révélée par les données d'échantillon, est suffisamment importante pour ne pas être due au hasard de l'échantillonnage ? En revanche, la portée correspond à l'applicabilité des résultats d'une étude à d'*autres populations ou contextes*. Par exemple, l'hypothèse d'une inégalité salariale de genre est-elle applicable en Afrique ? Il est plausible que les écarts de revenus, entre hommes et femmes, soient plus importants dans cette région en voie de développement qu'au Canada. La portée amène à réexaminer la littérature existante et ouvre potentiellement la voie à de futures recherches.

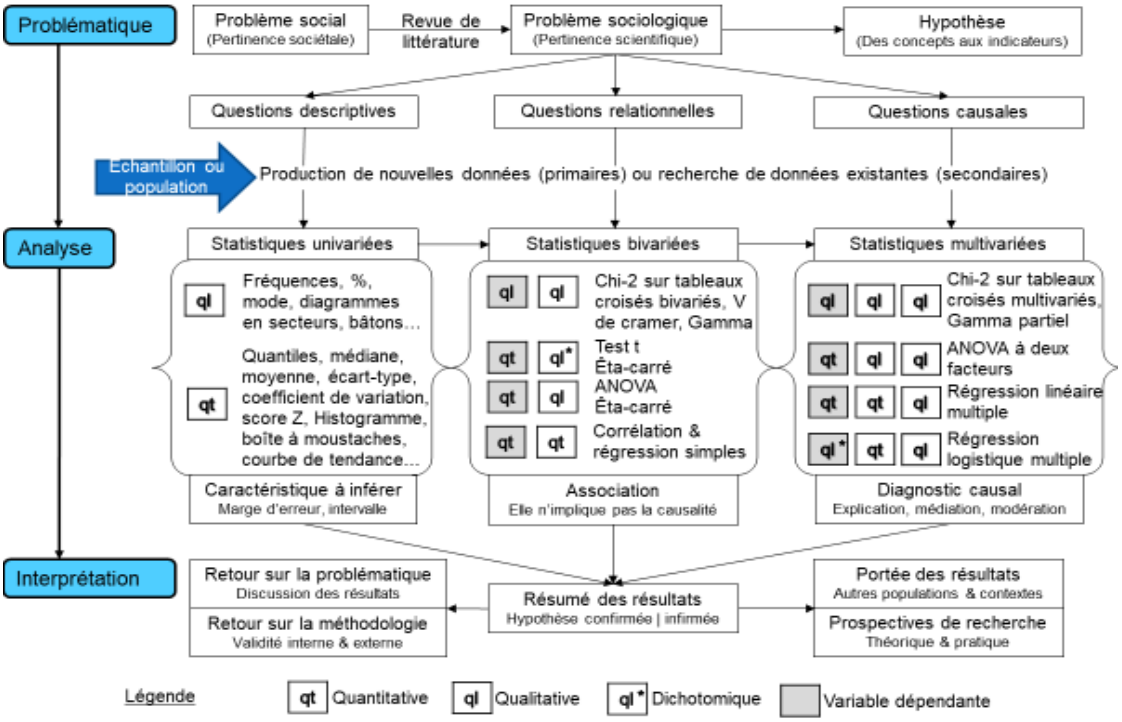
1.7. Recherche quantitative vs. Recherche qualitative

Pour bien saisir l'intérêt spécifique des méthodes quantitatives, il nous a paru important de les replacer dans le champ de la méthodologie de recherche en sciences sociales. À cet effet, les statistiques sont au cœur de la recherche quantitative. En tant qu'informations et outils d'analyse, elles facilitent notre compréhension du monde social. Leur usage s'inscrit au centre d'un va-et-vient incessant entre empirie et théorie, observations et hypothèses. Comme on peut le voir à travers le diagramme schématique ci-dessous, le raisonnement statistique transparaît dans toutes les étapes de ce processus cyclique.

Des questions de nature descriptive, relationnelle et causale peuvent être amalgamées dans une même étude quantitative, de façon à utiliser les statistiques uni-bi-multivariées. Ce que vous apprendrez à faire dans un cours plus avancé. Ce qu'il faut connaître dès à présent, c'est que pour un spécialiste des sciences sociales, savoir mener une recherche quantitative compte autant que la maitrise des statistiques elles-mêmes. Si utiles soient-elles, les statistiques ne peuvent pas se substituer à une problématisation rigoureuse, à une opérationnalisation des hypothèses, à la collecte de données probantes, et à une utilisation créative des connaissances théoriques et factuelles dans l'interprétation des résultats.

NE PAS COPIER

Figure 1.7. Place des statistiques dans le processus de recherche quantitative



La recherche quantitative présente des caractéristiques particulières qui la distinguent de la recherche qualitative. Pour s'en assurer, il suffit de les comparer en dressant quelques éléments de différenciation abordés en filigrane dans ce chapitre (tableau 1).

Bien qu'il y ait des différences indéniables qui séparent les méthodes quantitatives et qualitatives, elles ne s'excluent pas mutuellement. Une recherche est davantage convaincante lorsqu'elle combine ces deux méthodes. On parle alors de recherche mixte, laquelle autorise un portrait à la fois global et nuancé d'un phénomène. Il est très fécond de mener une *recherche mixte de type séquentiel explicatif* qui consiste à analyser des données quantitatives à l'aide des statistiques, puis à expliquer les résultats quantitatifs à l'aide de données qualitatives. Par exemple, une étude quantitative confirmant l'hypothèse de la relation entre le sexe et le revenu peut recourir à des focus groups, incluant des hommes et des femmes, pour tenter d'expliquer les inégalités salariales de genre. Également, Durkheim (1897) aurait pu recourir aux « motifs présumés des suicides », qu'il a sous-estimés, comme données qualitatives pour mieux expliquer les processus qui conduisent au suicide.

Tableau 1.2. Comparaison entre la recherche quantitative et la recherche qualitative

Items comparés	Recherche quantitative	Recherche qualitative
Origine épistémologique	Normative/positiviste (sciences sociales= sciences physico-naturelles)	Interprétative (Sciences de l'esprit ≠ Sciences physico-naturelles)
Visée scientifique	Description & explication des phénomènes sociaux	Exploration & compréhension des phénomènes sociaux
Rapport Sujet/Object	Extériorité (objectivité, objet vu comme une « chose »)	Intériorité (subjectivité, neutralité axiologique)
Raisonnement	Démarche déductive (des théories aux faits- logique de preuve)	Démarche inductive (des faits aux théories- logique de découverte)
Comportement humain	Régulier, prédictible	Dynamique, situationnel
Objet d'étude	Ampleur des faits (importance, tendances) & association, causalité entre des variables opérationnelles (régularités)	Déroulement des faits (vécu, perceptions) & correspondance entre des catégories (homologie de structure)
Procédures techniques	Rigoureuses, structurées & systématiques (expérimentation, enquête par questionnaire)	Souples, peu structurées & évolutives (entrevue, focus-group, observation participante)
Mesure	Précision (quantité), questions prédéterminées axés sur les résultats	Approximation (qualité), questions ouvertes axées sur les processus
Échantillonnage	Surtout probabiliste (simple, complexe)	Surtout non probabiliste (saturation, choix raisonné)
Échantillon	Larges unités statistiques, représentatives de la population	Cas restreints, crédibles
Type d'informations	Extensives (données numériques)	Intensives (verbatim, texte, vidéo)
Type d'analyses	Analyse statistique : Description statistique & inférence statistique	Analyse de contenu/thématique : Description narrative & contextuelle, comparaison continue
Résultats	Caractère nomothétique, généralisation	Caractère idiographique, transférabilité
But ultime	Réplicabilité, certitude, prédiction des phénomènes	Fiabilité, sens objectif, signification des phénomènes

Note : la liste des comparaisons n'est pas exhaustive.

NE PAS COPIER