Analyse quantitative en sciences humaines

Balthazar Charles

26 mai 2026

Table des matières

Première partie

Les préalables

Chapitre I

Utilité et limites de l'analyse quantitative

I.1 Les sciences humaines sont des sciences

Bien que les sciences humaines étudient des phénomènes complexes liés aux comportements, aux sociétés et aux cultures, elles partagent avec les sciences naturelles une approche rigoureuse et systématique de la connaissance. Ce qui confère aux sciences humaines leur caractère scientifique, c'est que la pratique des sciences humaines suit la méthode scientifique, présentée dans la figure I.1.

§1.1

L'utilisation de méthodes quantitatives en sciences humaines suit de près l'introduction des méthodes quantitatives en sciences en général. En effet, au XVIIe siècle, où émergent simultanément la méthode scientifique et la notion moderne d'État, la nécessité d'administrer les ressources de manière efficace pousse les gouvernements à collecter des données sur la population, le territoire, le commerce, etc. Bien que la collecte de données soit ancienne (recensements en Mésopotamie, Égypte, Chine antique), c'est à cette époque que se développe

une approche systématique de la collecte et de l'analyse des données pour comprendre et gérer les sociétés.

Exemple I.1.0

Au Québec, le premier recensement officiel a eu lieu en 1666 sous l'administration de Jean Talon, intendant de la Nouvelle-France. Ce recensement visait à collecter des données démographiques et économiques pour mieux administrer la colonie. Il a permis de recenser environ 3 215 habitants, fournissant des informations précieuses sur la population, les familles, les métiers et les ressources disponibles. Ce recensement est un exemple précoce de l'utilisation de méthodes quantitatives pour comprendre et gérer une société.

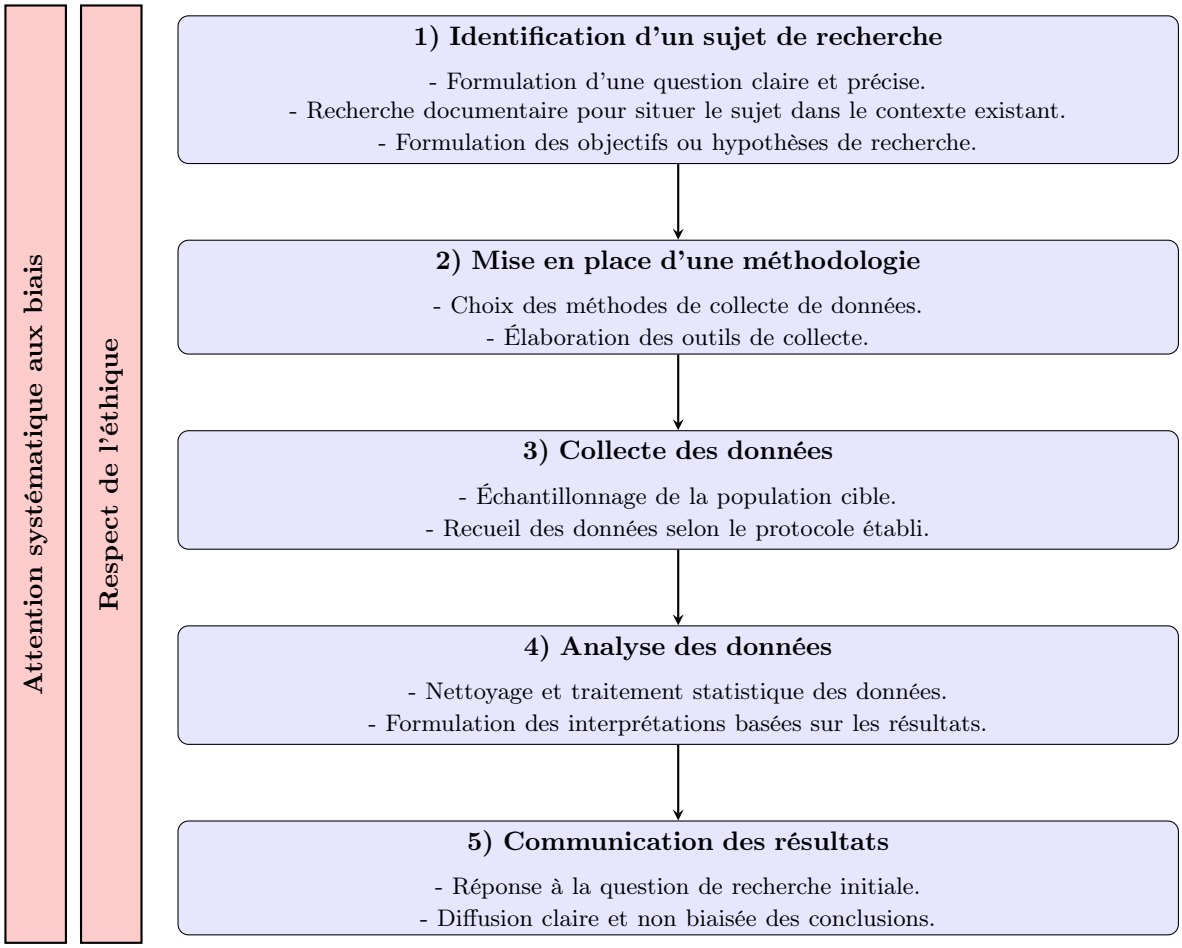


FIGURE I.1 – Structure de la méthode scientifique.

Cette tendance à la quantification ne s'est pas démentie depuis, entre recensements, sondages, nécessités logistiques, etc. De nos jours, impulsée par le succès des sciences dans l'amé-

lioration de la qualité de vie et dopée par l'avènement des outils informatiques, l'analyse quantitative est devenue un outil incontournable dans tous les domaines. En plus de permettre aux chercheurs de mener des études rigoureuses et reproductibles, les méthodes quantitatives appliquées spécifiquement aux réalités humaines sont un support majeur de la prise de décision politique, économique et sociale.

Exemple I.1.1

En étudiant la diversité génétique chez les graminées, par des méthodes quantitatives, des archéologues ont pu prouver qu'avant l'arrivée du maïs en Amérique du Nord, les populations autochtones cultivaient une plante appelée "chénopode" et avaient domestiqué le tournesol (entre autres plantes). [a]

a. *Eastern North America as an independent center of plant domestication*, par Bruce Smith, 2006

Exemple I.1.2

Un urbaniste montréalais veut défendre l'installation de pistes cyclables. Il décide de citer l'article *The cars are going to be alright : Examining micromobility infrastructure space allocation and potential improvement scenarios in Montréal* par Daniel Romm, *et al.* de l'université McGill, publié en 2025. Dans cette étude, les chercheurs introduisent l'indicateur quantitatif "d'allocation égalitaire d'infrastructure" et montrent que, selon cette mesure, l'espace alloué aux cyclistes est moindre que celui alloué aux voitures.

Bien que les sciences humaines en elles-mêmes ne soient pas **normatives** (elles ne disent pas ce qu'il faut faire), la pratique de cette méthodologie scientifique en sciences humaines est souvent motivée par le besoin de données empiriques pour éclairer la prise de décision dans des domaines variés tels que la santé publique, l'éducation, la politique sociale, etc.

Dans ce cadre de décision, l'analyse quantitative en sciences naturelles et humaines fournit des données que l'on veut **objectives** pour justifier des choix ayant souvent comme visée une efficacité de l'allocation des ressources (dans une entreprise par exemple, mais aussi dans le secteur public : étant donné un budget fixé, comment l'utiliser au mieux pour maximiser le bien-être de la population ?). Cependant, il est crucial de reconnaître que les données quantitatives, bien qu'objectives dans leur collecte et analyse, sont interprétées par des humains qui apportent leurs propres perspectives, valeurs et biais. Ainsi, l'objectivité des données ne garantit pas une objectivité dans la prise de décision.

I.2 Qu'est-ce que l'analyse quantitative ?

L'analyse quantitative est une approche méthodologique qui utilise des données numériques et des techniques statistiques pour comprendre des phénomènes sociaux, psychologiques et

humains et qui vise à les mesurer et les chiffrer. Elle est une des manifestations de la méthode scientifique. Ce n'est pas la seule : les méthodes qualitatives ont également une place centrale en sciences humaines. Il faut noter que l'analyse quantitative est une méthodologie, et peut donc être appliquée à toutes sortes de questions de recherche et de données, y compris des données qualitatives.

Exemple I.2.0

Après les caisses dans un supermarché, une borne avec trois boutons (satisfait, neutre, insatisfait) permet de recueillir des avis qualitatifs sur l'expérience d'achat. Après un traitement quantitatif de ces données, on peut par exemple faire l'affirmation suivante : "85% des clients se disent satisfaits de leur expérience d'achat".

Définition I.2.1 : Analyse quantitative

L'**analyse quantitative** est une méthode de recherche qui s'appuie sur la collecte et l'analyse numériques de données pour étudier des phénomènes, tester des hypothèses et établir des relations entre variables.

Pour le moment, ces mots ne nous disent pas grand-chose. L'objectif de la première partie de ce cours est de se familiariser avec les concepts, outils et méthodes de l'analyse quantitative en sciences humaines.

Cette approche repose sur plusieurs principes fondamentaux :

- **Mesure objective** : Utilisation d'instruments standardisés et de critères précis
- **Analyse statistique** : Application de méthodes mathématiques pour décrire et interpréter les données
- **Généralisation** : Possibilité d'étendre les conclusions d'un échantillon à une population plus large
- **Reproductibilité** : Les résultats peuvent être vérifiés et répliqués par d'autres chercheurs

D'une façon générale, tous les praticiens des sciences humaines vont, d'une façon ou d'une autre, être confrontés à l'analyse quantitative, parfois comme auteurs qui utilisent les outils fournis par les méthodes quantitatives pour mener leur recherche, souvent comme lecteurs qui doivent comprendre et interpréter des résultats quantitatifs présentés dans des articles scientifiques, rapports, etc. Plus largement que pour la recherche scientifique, les résultats d'analyses quantitatives sont une des sources majeures d'exposition du public aux sciences mathématiques, après la finance. En effet, on trouve très régulièrement dans les médias des sondages d'opinion, des statistiques sur des phénomènes sociaux, économiques ou sanitaires, etc. Comprendre les bases de l'analyse quantitative est donc un atout important pour être un citoyen informé et critique, y compris pour les gens dont l'activité principale est aussi éloignée

que possible des mathématiques.

I.3 Quand choisir l'analyse quantitative ?

Les analyses quantitatives et qualitatives ne sont en aucun cas exclusives l'une de l'autre et un même projet de recherche peut combiner les deux. Cependant, certaines situations ou objectifs de recherche se prêtent plus naturellement que d'autres à une approche quantitative.

Analyse quantitative	**Analyse qualitative**
État de la recherche sur le sujet	
Sujet documenté, pour lequel on a une vue d'ensemble.	Tout type de sujet, y compris peu exploré, nécessitant une compréhension approfondie.
Exemple : Mesurer l'impact d'une politique de santé : on dispose déjà de nombreuses manières de le mesurer.	**Exemple :** La polarisation politique entre les genres chez les jeunes adultes augmente : on veut en comprendre les raisons.
Objectifs de recherche	
Mesurer un phénomène, tester des hypothèses, établir des relations de cause à effet. Obtenir des données chiffrées, comparables et généralisables. Questions dont on peut estimer la réponse à l'avance.	Comprendre un phénomène difficilement chiffrable, explorer des expériences subjectives, générer des hypothèses. Obtenir une information riche, protéiforme, dont on ne peut précisément prévoir la nature.
Exemple : on peut vouloir mesurer l'impact d'une nouvelle méthode pédagogique sur les résultats scolaires : on recueille les notes des étudiants avant et après l'introduction de la méthode.	**Exemple :** on peut vouloir comprendre les raisons du rejet d'une réforme : on recueille les avis des intéressés par écrit. On obtient un corps de textes dont il faut extraire les raisons exprimées
Type de données	
Données structurées, souvent numériques.	Chaque donnée est souvent complexe et riche en informations qualitatives.
Exemples : Réponse à un questionnaire à choix multiples, relevés de notes d'examen, données démographiques...	**Exemples :** des lettres historiques, des documents ethnologiques filmés, des suggestions d'utilisateurs d'un service
Volume de données	
Grand volume, jusqu'à plusieurs milliards	Faible volume, jusqu'à quelques centaines

Généralement parlant, l'analyse quantitative est plus restrictive que l'analyse qualitative sur les questions de recherches qu'elle peut aborder et les données qu'elle peut traiter. En contrepartie, les résultats obtenus sont plus précis, plus clairement interprétables et souvent généralisables.

Évidemment, les différences entre méthodes quantitatives et qualitatives sont plus nombreuses et plus nuancées que ce bref résumé. De plus, les conclusions tirées d'une approche peuvent nourrir la recherche menée dans l'autre : par une approche quantitative, on peut montrer que certaines variations dans la population au cours du temps sont significatives, ce qui peut motiver une étude qualitative pour en comprendre les raisons. Inversement, une étude qualitative peut éclairer les interprétations possibles d'une analyse quantitative et permettre de tirer des conclusions plus riches que les simples chiffres.

Exemple I.3.0

La chercheuse montréalaise en géographie urbaine, la Pr. Damaris Rose, propose en 1984 dans son article *Rethinking gentrification : beyond the uneven development of marxist urban theory* une analyse qualitative des processus qui mènent à la gentrification d'un quartier (dans le cadre des villes nord-américaines). Son analyse suggère en particulier qu'on peut s'attendre à des compositions sociales différentes au cours du processus. Dans des travaux ultérieurs, la Pr. Rose et d'autres chercheurs ont mesuré quantitativement ces compositions, par exemple dans l'article *Economic restructuring and the diversification of gentrification in the 1980s : a view from a marginal metropolis* de 1996, dans lequel les mesures quantitatives des changements démographiques dans plusieurs quartiers de Montréal sont analysées pour tester les hypothèses formulées dans l'article de 1984.

I.4 Importance en sciences humaines

§1.3

L'analyse quantitative en sciences humaines est utile pour les mêmes raisons qu'elle l'est en sciences naturelles : elle permet de structurer la collecte et l'analyse des données et de produire des résultats reproductibles. Bien qu'il soit difficile [1] d'atteindre une objectivité totale en sciences humaines, les méthodes quantitatives permettent de créer des instruments de mesure des réalités humaines standardisés et réutilisables.

Exemple I.4.0

On peut argumenter que les tests de QI ne mesurent pas réellement "l'intelligence" dans son ensemble, mais à défaut de pouvoir accéder à une mesure "parfaite" (dont

1. Et, pourrait-on argumenter, impossible et dans certains cas, pas nécessairement souhaitable.

l'existence même est loin d'être garantie), ces tests fournissent un moyen standardisé d'évaluer certains aspects cognitifs. Grâce à cette standardisation, les résultats peuvent être comparés entre individus et groupes, et utilisés pour des recherches en psychologie cognitive, en éducation, etc.

En chiffrant différents aspects des réalités humaines, l'analyse quantitative permet de :

- **décrire rigoureusement** des phénomènes complexes et possiblement de grande échelle. Plutôt que de décrire une population comme "très pauvre", on peut préciser que son revenu est inférieur à 50% du revenu médian national. Plutôt que d'examiner des comportements individuels, on peut analyser des tendances dans une population de plusieurs millions de personnes.
- ce faisant, elle permet de **comparer des groupes** qualitativement différents : être pauvre en Suisse n'a pas le même sens qu'être pauvre au Venezuela. Cependant, en utilisant des mesures relatives, on peut comparer la pauvreté entre ces deux pays. Le chiffrage des phénomènes humains permet également aux chercheurs de répliquer des études et mesures et d'obtenir des données comparables entre elles.
- **généraliser les résultats** d'une étude à une population plus large : en utilisant des échantillons représentatifs et des méthodes statistiques appropriées, on peut tirer des conclusions sur une population entière à partir d'un sous-ensemble de cette population.
- **établir des liens** entre des phénomènes, tester des hypothèses et faire des prévisions : en observant quantitativement la relation entre les politiques familiales et les taux de natalité dans des pays comparables au Canada, on peut faire des prédictions sur l'impact de telles politiques au Canada.
- **aider à la prise de décision**, une fois appliquée à des problématiques concrètes.

Donnons des exemples concrets d'application dans différents domaines des sciences humaines. Cette liste n'est pas exhaustive, mais illustre la diversité des applications de l'analyse quantitative en sciences humaines.

Domaine	Applications de l'analyse quantitative
Administration	Création de modèles budgétaires basés sur des données historiques pour optimiser l'allocation des ressources. Projection de l'impact économique d'une nouvelle infrastructure ou de l'accueil d'un événement international.

Anthropologie et Archéologie	Analyse de la diversité génétique des populations anciennes pour comprendre les migrations humaines. Utilisation de la datation par le carbone 14 pour établir des chronologies précises. Prédiction des sites archéologiques potentiels à l'aide de modèles spatiaux.
Démographie	Création de recensements pour mesurer la population et ses caractéristiques. Modélisation des tendances démographiques pour prévoir les besoins futurs en infrastructures, services sociaux, etc. Calculs de nombreux indicateurs sociaux-économiques (taux de natalité, mortalité, migration, etc.) informant des politiques publiques.
Économie	Analyse des données de marché pour comprendre les comportements des consommateurs. Modélisation économétrique pour prévoir les impacts des politiques fiscales ou monétaires. Évaluation quantitative des risques financiers. Calculs des taux d'intérêt, primes d'assurances, etc.
Géographie	Utilisation de systèmes d'information géographique (SIG) pour analyser les données spatiales. Modélisation des phénomènes environnementaux (changement climatique, urbanisation, etc.) et leur impact sur les populations humaines.
Histoire	Analyse statistique des données historiques (recensements, registres économiques, etc.) pour comprendre les tendances sociales et économiques. Utilisation de méthodes quantitatives pour étudier les réseaux sociaux de nos jours ou les dynamiques de pouvoir dans les sociétés passées.
Psychologie	Utilisation de tests standardisés pour mesurer les traits de personnalité, les capacités cognitives, etc. Analyse statistique des données expérimentales pour comprendre les comportements humains et les processus mentaux.
Science politique	Analyse des données électorales pour comprendre les comportements de vote. Modélisation des opinions publiques à l'aide d'enquêtes quantitatives. Évaluation de l'impact des politiques publiques à l'aide de méthodes statistiques.

Sociologie	Études quantitatives des comportements sociaux, des structures familiales, des inégalités sociales, etc. Utilisation de grandes bases de données pour analyser les tendances sociales à long terme.

I.5 Limites

I.5.1 Une approche descriptive, pas normative.

Fondamentalement, la science est descriptive : elle cherche à comprendre, expliquer, modéliser le monde tel qu'il est, pas tel qu'il devrait être. C'est le cas également en sciences humaines. L'analyse quantitative permet de mesurer des phénomènes humains, mais ne prescrit pas d'action à prendre. Par exemple, une étude peut montrer un lien entre certaines politiques sociales et une réduction de la pauvreté, une autre peut mesurer le niveau de pauvreté dans la population, mais aucune ne dit qu'il faut adopter ces politiques sociales.

Bien que l'analyse quantitative et ses conclusions soient des outils puissants pour indiquer la façon d'agir en poursuite d'un objectif, la définition de cet objectif et l'acceptabilité des actions à prendre relèvent du domaine politique, éthique et moral, qui dépasse le cadre de la science. Comme le dit David Hume dans son *Traité de la nature humaine* (1739) : "On ne peut jamais déduire un *il faut* d'un *il est*". En d'autres termes, les faits seuls ne peuvent pas dicter ce qui devrait être fait.

Notons cependant que dans vos vies professionnelles, si vous travaillez dans l'administration, la santé publique, le marketing..., il sera fréquent que les objectifs soient imposés à l'avance et que votre rôle soit de fournir et analyser des données pour éclairer la prise de décision.

I.5.2 Des réalités parfois trop complexes

Malgré ses nombreux avantages, l'analyse quantitative a des limites inhérentes, en particulier en sciences humaines. Les phénomènes humains sont souvent complexes, multifactoriels et contextuels, ce qui peut rendre difficile leur quantification précise. Par exemple, des concepts comme le bonheur, la justice ou la culture sont difficiles à définir, et donc à mesurer de manière objective et standardisée. Dans certains cas, une approche qualitative est incontournable pour comprendre pleinement ces phénomènes. Par essence, les méthodes quantitatives simplifient la réalité pour la rendre mesurable, ce qui peut entraîner une perte d'informations importantes. Il faut donc, lorsque l'on revient des chiffres vers la réalité lors de l'interprétation des résultats, garder à l'esprit cette simplification : il est possible que des aspects cruciaux du phénomène étudié aient été négligés ou mal représentés.

§1.4

Si malgré tout on essaie de réduire ces phénomènes à des chiffres (ce qui peut se justifier),

on prend le risque que la définition personnelle du chercheur ou des biais culturels influencent la manière dont les phénomènes sont mesurés et interprétés. Par exemple, une enquête sur la satisfaction au travail peut être influencée par des facteurs culturels qui affectent la manière dont les individus expriment leur satisfaction ou insatisfaction ou sur la perception de ce qu'est le niveau d'implication "normal".

I.5.3 Considérations éthiques.

Éthique de la recherche Les mêmes considérations d'éthique de la recherche qui s'appliquent dans toutes les sciences s'appliquent en particulier en sciences humaines. Il est crucial d'être aussi transparent que possible sur les méthodes utilisées, les limites des données et les interprétations possibles des résultats. De ce point de vue, la méconduite scientifique peut prendre de nombreuses formes : falsification de données, plagiat, biais dans la collecte ou l'analyse des données, analyse "dirigée" pour obtenir des résultats souhaités, etc.

§1.5

Exemple I.5.0

En mai 2025, le département des Sciences du Comportement Humain de l'université Harvard a renvoyé Francesca Gino, dont les recherches portaient sur "le comportement éthique et l'honnêteté", pour avoir falsifié, voire inventé, des données dans plusieurs articles (au moins 4) entre 2012 et 2021.

Ce genre de comportement nuit non seulement à la crédibilité du chercheur, mais aussi à la confiance du public dans la recherche scientifique en général. Sur la base d'articles publiés illégitimement, des politiques publiques peuvent être mises en place, des ressources allouées, des vies humaines impactées. D'autres recherches sont parfois entreprises pour tenter d'étudier un phénomène parfois inventé, ce qui gaspille du temps, de l'argent et de la crédibilité scientifique.

Exemple I.5.1

En 1998, Andrew Wakefield a publié une étude dans *The Lancet* portant sur 12 enfants et suggérant un lien entre le vaccin ROR (rougeole-oreillons-rubéole) et l'autisme. Cette étude a été largement discréditée par la suite en raison de graves erreurs méthodologiques, de conflits d'intérêts non divulgués, de manipulations des données et de brèches éthiques dans le traitement des patients. Cependant, l'impact de cette publication a été considérable, contribuant à une baisse de la vaccination et à des épidémies de rougeole dans plusieurs pays. L'étude a été menée au Royaume-Uni où elle a eu un impact important. Des décennies plus tard, cela continue à avoir des conséquences : selon les statistiques du gouvernement britannique, 10 morts causées par la rougeole auraient pu être évitées par la vaccination depuis 2010. De nombreuses études ultérieures ont infirmé

tout lien entre le vaccin ROR et l'autisme. Par exemple, une étude par Madsen *et al.* en 2002 [a], portant sur 537,303 enfants, n'a trouvé aucune association entre la vaccination ROR et le développement de l'autisme.

a. A population-based study of [MMR] vaccination and autism, New England Journal of Medicine.

Spécificité des sciences humaines Dans le cas des sciences humaines, des considérations supplémentaires sont à prendre en compte, en particulier en ce qui concerne d'une part les sujets de la recherche et d'autre part l'interprétation des résultats. §1.5.1

Concernant les sujets de recherche, il s'agit dans une majorité des cas directement d'humains ou de groupes humains. Il est donc crucial de respecter leur dignité, leur vie privée et leurs droits tout au long du processus de recherche. Cela inclut l'obtention d'un consentement éclairé, la protection des données personnelles et la minimisation des risques pour les participants. Dans certains cas, la question du consentement est particulièrement difficile : par exemple, un archéologue peut être amené à étudier des restes humains anciens ou un historien peut vouloir analyser des documents personnels. Dans ces cas, il est important de respecter les normes éthiques en vigueur et de consulter les communautés concernées lorsque cela est approprié. Les sciences humaines touchent à tous les aspects de la vie, y compris des sujets difficiles comme la pauvreté, la maladie, la violence, la discrimination, etc. Il est crucial de traiter ces sujets avec sensibilité et respect, en évitant de stigmatiser ou de marginaliser les groupes étudiés.

Exemple I.5.2

L'utilisation de nos traces numériques (données de navigation web, données de réseaux sociaux, etc.) par des entreprises privées dans des buts commerciaux ou par des gouvernements à des fins de surveillance soulève des questions éthiques importantes. Par exemple, en 2018, le scandale Cambridge Analytica a révélé que les données personnelles de millions d'utilisateurs de Facebook avaient été collectées sans leur consentement pour influencer des campagnes politiques. Cet incident a mis en lumière la nécessité de réglementations strictes sur la collecte et l'utilisation des données personnelles, ainsi que l'importance du consentement éclairé dans la recherche impliquant des données numériques.

En ce qui concerne l'interprétation des résultats, il est important de reconnaître et de rendre explicites les limites des données et des méthodes utilisées. Les chercheurs doivent être transparents sur les incertitudes et les biais potentiels, et éviter de tirer des conclusions hâtives ou non fondées. De plus, il est essentiel de considérer le contexte culturel, social et historique des données, afin d'éviter les généralisations inappropriées ou les interprétations erronées.

Exemple I.5.3

À l'époque du "racisme scientifique" au XIX^{ème} siècle, des mesures anthropométriques étaient utilisées pour justifier des hiérarchies raciales. En particulier, des mesures de volumes crâniens étaient interprétées comme des indicateurs de supériorité intellectuelle, et de nombreux chiffres de l'époque, remesurés sur des spécimens conservés dans les musées, montrent que les données avaient été manipulées pour correspondre aux préjugés raciaux des chercheurs.

Il est de la responsabilité du chercheur de se tenir à l'affût de mauvaises données ou d'interprétations dépassées et/ou biaisées, et de les éviter ou de les corriger selon le cas. De la même façon, il faut garder à l'esprit que les données quantitatives ont un grand pouvoir de persuasion, et que les chiffres peuvent être utilisés pour manipuler l'opinion publique ou justifier des politiques controversées. Tout "objectifs" que soient les chiffres, il est possible de les faire mentir : nous en verrons des exemples lorsque nous parlerons de la présentation et de l'interprétation des résultats. Les chercheurs doivent être conscients de cette réalité et s'efforcer de communiquer leurs résultats de manière honnête et transparente. Cela est d'autant plus vrai quand les questions traitées touchent à des enjeux sociaux sensibles, comme les inégalités, la discrimination, la santé publique, etc. Enfin, et bien que ce dernier point soit vrai dans toutes les sciences, les sciences humaines sont souvent particulièrement proches de leur propre application. Par exemple, un sociologue étudiant les politiques publiques peut être amené à conseiller des décideurs politiques sur la base de ses recherches. Dans ces situations, il est crucial de maintenir une séparation claire entre la recherche scientifique et les intérêts politiques ou économiques, afin de préserver l'intégrité et l'objectivité de la recherche, quelle que soit la préférence politique personnelle du chercheur.

En résumé, voici une liste non exhaustive de problèmes éthiques qui peuvent survenir à différentes étapes du processus de recherche en sciences humaines, et qui peuvent affecter la qualité des données et la validité des conclusions :

— **Problème impactant les sujets de l'étude :** Violation de la vie privée des participants et utilisation de données sensibles sans précautions adéquates, collecte de données sans consentement éclairé, exposition des participants à des risques physiques ou psychologiques, exploitation de populations vulnérables, etc.

— **Problèmes impactant directement la qualité des données :** Manipulation ou falsification de données, utilisation de données obsolètes ou non représentatives, etc.

— **Problèmes impactant directement la validité des conclusions :** Interprétation biaisée des résultats pour soutenir un agenda particulier, incluant la validation de préjudices, la stigmatisation ou marginalisation des groupes étudiés, le gain personnel, conflits d'intérêts non divulgués, publication de résultats non vérifiés ou non reproductibles, etc.

Évidemment, un problème en amont peut entraîner des problèmes en aval : par exemple, la collecte de données sans consentement éclairé peut conduire à des données biaisées, ce qui peut à son tour conduire à des conclusions erronées ou biaisées. De même, l'interprétation biaisée des résultats peut conduire à des politiques publiques inappropriées ou injustes. Il est donc crucial de maintenir des normes éthiques élevées à chaque étape du processus de recherche pour garantir la qualité et l'intégrité de la recherche en sciences humaines.

Résumé du chapitre

Concepts clés

Concept	Définition
Analyse quantitative	Méthode d'étude utilisant des données numériques et techniques statistiques pour comprendre et mesurer des phénomènes
Objectivité	Qualité d'une mesure standardisée (données), pas des décisions (normatives)
Généralisation	Possibilité d'étendre les conclusions d'un échantillon à une population plus large
Reproductibilité	Les résultats peuvent être vérifiés et répliqués par d'autres chercheurs

Quand utiliser l'analyse quantitative ?

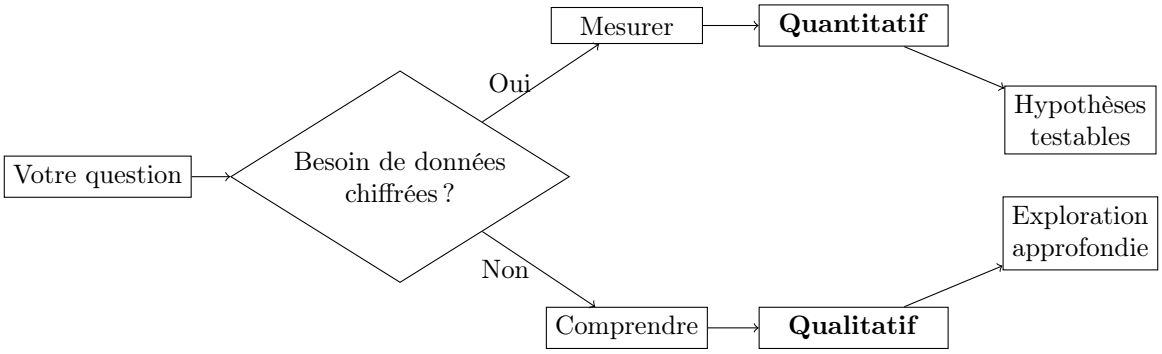


FIGURE I.2 – Décision entre approches quantitative et qualitative

Forces et limites

Forces :

- Mesure objective et standardisée
- Comparaison entre groupes possibles
- Généralisation valide
- Décisions éclairées par données
- Reproductibilité assurée

Limites :

- Réalité humaine complexe
- Perte d'information contextuelle
- Pas de prescriptions (normatif)
- Biais d'interprétation possibles
- Considérations éthiques importantes

Principes éthiques fondamentaux

1. **Consentement éclairé** : Les participants doivent comprendre et accepter l'étude
2. **Confidentialité** : Protéger les données personnelles
3. **Transparence** : Documenter les méthodes et limites
4. **Intégrité** : Éviter la falsification et le biais
5. **Respect** : Traiter les groupes sans stigmatisation

Chapitre II

Démarche méthodologique

On l'a dit, les méthodes quantitatives sont une manifestation de la méthode scientifique. La phase d'analyse elle-même (et dans une moindre mesure, la phase de communication) constitue l'essentiel de ce cours, mais on voit dans cette section les autres étapes de la démarche scientifique en analyse quantitative.

II.1 Identification d'un sujet de recherche

La formulation d'une question de recherche claire et précise est la première étape cruciale de toute étude quantitative. Une bonne question de recherche doit être : §2.1.1

— **Spécifique** : Éviter les questions trop larges ou vagues,

— **Mesurable** : Les concepts doivent pouvoir être quantifiés,
— **Réaliste** : Faisable avec les ressources disponibles,
— **Pertinente** : En lien avec les connaissances existantes,
— **Originale** : Apportant une contribution nouvelle au domaine.

Exemple II.1.0 : Question de recherche

Question vague : "Les étudiants sont-ils stressés ?"
Question précise : "Quel est le niveau moyen de stress perçu (mesuré sur une échelle de 1 à 10) chez les étudiants de première année de cégep pendant la période d'examens finaux, et ce niveau diffère-t-il selon le programme d'études ?"
La deuxième formulation spécifie la population (étudiants de première année de cégep), la variable à mesurer (stress sur échelle 1-10), le moment (période d'examens), et une variable comparative (programme d'études).

On peut, au premier abord, distinguer deux types de questions de recherche : celles qui cherchent simplement à faire une observation, mesurer un concept, pour lesquelles on a un *objectif de recherche*, et celles qui cherchent à établir des relations entre plusieurs concepts, ou vérifier une "explication", pour lesquelles on établit une *hypothèse*, l'objectif étant de vérifier ou invalider cette hypothèse.

Définition II.1.1 : Hypothèse

Une **hypothèse** est une proposition *testable* (donc *précise*) qui suggère une relation entre deux ou plusieurs variables. Elle est formulée *avant* la collecte des données et sert de base à l'analyse statistique.

Dans le cadre de la méthode scientifique, et pour un but de recherche explicatif, la formulation de l'hypothèse *à l'avance* est une étape cruciale. D'une part, le contenu de l'hypothèse informe la construction de l'expérience et la collecte des données. D'autre part, formuler l'hypothèse à l'avance permet au chercheur de ne pas être influencé par les données lors de la formulation de l'hypothèse, ce qui pourrait biaiser l'analyse. En effet, il est toujours facile de formuler une hypothèse en ayant vu les données, mais ce qui permet de décider de la validité d'une hypothèse, c'est sa capacité à prédire des résultats *avant* de les voir. Nous verrons plus tard dans ce cours comment tester formellement des hypothèses à l'aide de tests statistiques. Nous verrons également comment sans formuler d'hypothèse à l'avance, on prend le risque de "trouver" des relations qui n'existent pas réellement dans les données.

En fait, quand on teste une hypothèse, on compare deux hypothèses alternatives : l'hypothèse nulle (notée H_0) et l'hypothèse alternative (souvent notée H_a). L'hypothèse nulle est généralement une affirmation de "non-effet" ou "non-relation", tandis que l'hypothèse alter-

native est ce que le chercheur cherche à montrer. Par exemple, si l'on veut tester si un nouveau médicament est efficace, l'hypothèse nulle pourrait être "le médicament n'a pas d'effet sur la maladie", tandis que l'hypothèse alternative serait "le médicament améliore la condition des patients". Dans certains cas, quand des connaissances existent déjà sur le sujet, l'hypothèse nulle peut être plus spécifique que "pas de relation" et l'hypothèse alternative en constitue une version qu'on espère plus précise. Par exemple, on peut déjà savoir qu'il existe une relation entre deux variables, mais on veut tester si cette relation est positive ou négative.

II.2 Opérationnalisation et outils de collecte des données

II.2.1 Opérationnalisation

L'opérationnalisation consiste à transformer des concepts abstraits en variables concrètes et mesurables. C'est le processus qui permet de passer de l'idée théorique à la mesure empirique. §2.1

Par exemple, pour décider des directions à suivre dans une politique de santé publique, on pourrait avoir envie de mesurer "l'accès aux soins". Tel quel, ce concept est trop vague pour pouvoir être mesuré. Il faut donc *choisir* quel sens quantitatif on donne à ce concept. On peut choisir de l'opérationnaliser comme :

— le "nombre de centres de santé",
— la "fréquence des visites médicales",
— le "temps de trajet pour se rendre à l'hôpital le plus proche", etc.

On peut également choisir de combiner plusieurs de ces aspects en un indice composite. Cependant, chacun des points précédents n'est pas encore assez précis : la "fréquence des visites médicales" doit être précisée. Quelle période de temps ? Quel type de visites médicales retient-on ? Veut-on mesurer la (sur)charge de travail des médecins, auquel cas on compte les visites par médecin, ou la facilité à voir un médecin, auquel cas on compte les visites par habitant ? Les mêmes questions se posent pour les autres aspects proposés. Enfin, si on choisit de créer un indice composite, comment combiner ces différents aspects ? Toutes ces décisions doivent être prises *et justifiées* lors de l'opérationnalisation.

Définition II.2.0 : Opérationnalisation

L'**opérationnalisation** est le processus de décision et de définition des aspects concrets qui permettront de mesurer ou d'observer un concept théorique.

Une opérationnalisation correcte d'une question en un protocole de mesure vise deux objectifs fondamentaux : §2.4

— **Validité** : La mesure doit refléter fidèlement le concept qu'elle est censée représenter.
— **Fidélité** : La mesure doit être cohérente et reproductible dans le temps et entre différents observateurs.

§2.4.1 **Validité** Dans certains cas, il est très facile d'obtenir une mesure valide : la taille ou l'âge d'une personne, son revenu annuel, la quantité d'outils retrouvés sur des sites archéologiques, etc, peuvent être mesurés directement. Le plus souvent, on cherche en science humaine à discuter des réalités humaines plus compliquées : par exemple, l'effet de la beauté sur la réussite professionnelle [1], la relation entre salaire et confiance pour le futur, les retombées économiques du REM... La chose que l'on veut mesurer dans ces cas n'est pas directement accessible : on mesure à la place des quantités que l'on juge indicatives de la quantité d'intérêt. Même si la validité de la mesure est rarement parfaite, car les concepts étudiés sont souvent complexes, il est important de savoir justifier l'effet des choix d'opérationnalisation sur la validité de la mesure.

Exemple II.2.1

Je suis votre professeur, et je veux évaluer votre compréhension du cours. Je ne peux pas mesurer directement votre "compréhension", mais je peux mesurer vos résultats aux examens, que je considère comme un indicateur de votre compréhension. Cependant, si je choisis de mettre des exercices de calcul différentiel dans l'examen d'analyse quantitative, la validité de ma mesure sera faible, car la capacité à faire du calcul différentiel n'est pas un bon indicateur de votre compréhension de l'analyse quantitative.
Dans le cas d'un cours, on considère souvent par défaut que l'enseignant sait préparer une évaluation pertinente. Cependant, il arrive que les élèves se plaignent de la difficulté d'un examen, ou qu'il contienne un sujet non vu en classe. Dans ce cas, les élèves remettent en question la validité de l'indicateur "note d'examen" pour mesurer leur compréhension du cours.

§2.4.2 **Fidélité** Il est naturel de vouloir qu'un instrument de mesure, appliqué deux fois dans les mêmes conditions, donne le même résultat. La fidélité d'un instrument de mesure est une mesure de sa cohérence et permet de s'assurer que les différentes mesures sont comparables entre elles. Pour cela, il est souvent nécessaire d'être précis et explicite dans la définition des procédures de mesure.

Exemple II.2.2

(Exemple fictif) On veut mesurer l'optimisme concernant l'économie mondiale chez les jeunes adultes. On crée un questionnaire permettant de mesurer ce concept. On communique ensuite ce questionnaire le 21 décembre à 1000 jeunes adultes canadiens et 1000 jeunes adultes australiens. On répète l'expérience le 21 juin de l'année suivante, en utilisant le même questionnaire et la même procédure. On note que les Canadiens sont

1. Cherchez "effet de halo".

en moyenne plus confiants lors de la seconde mesure, tandis que les Australiens sont en moyenne moins confiants. Le fait que les deux groupes aient des évolutions opposées alors qu'ils sont exposés aux mêmes événements économiques mondiaux suggère que le questionnaire utilisé n'est pas fidèle : il est sensible à des facteurs externes non contrôlés (par exemple, la saison, le climat, les événements locaux, etc.) et ne permet pas de mesurer de manière cohérente l'optimisme économique. Peut-être que l'été rend optimiste et que les variations opposées viennent de la différence de saisons ?

Il faut faire attention à concevoir un protocole de mesure qui, autant que possible, ne dépend que du concept que l'on veut mesurer et de la population (ou échantillon) étudiée, et pas d'autres facteurs externes. Cela mène par exemple à vérifier que les mesures effectuées par différents observateurs donnent des résultats similaires (fidélité inter-juges), ou que les mesures répétées dans le temps sont cohérentes (fidélité test-retest).

Exemple II.2.3

Dans l'examen du ministère, si les notes mises par un correcteur sont systématiquement plus élevées que celles d'un autre, la fidélité inter-juges est faible. Pour améliorer cette fidélité, on peut donner des consignes de correction précises aux correcteurs et appliquer une *modération* aux notes d'un groupe de copies.

Être fidèle ne garantit pas d'être valide, et inversement. Par exemple, se peser avec une balance dont la tare n'est pas faite donne des mesures fidèles, mais pas valides. De même, les mesures phrénologiques (mesure des bosses sur le crâne) étaient considérées comme fidèles (plusieurs observateurs mesuraient les mêmes bosses de la même manière), mais pas valides (elles ne mesuraient pas réellement les traits de personnalité).

La validité d'un test est souvent plus difficile à établir que sa fidélité, pour laquelle on peut appliquer "réflexivement" les méthodes de l'analyse quantitative : nous y reviendrons.

II.2.2 Comment opérationnaliser un concept ?

La réponse dépend fortement de la question de recherche, du concept à mesurer, mais aussi des moyens de l'étude et des restrictions qui l'encadrent. Cependant, en général, on peut partir du principe qu'*opérationnaliser, c'est décomposer*. Si la question de recherche s'intéresse à un domaine très restreint, on peut souvent transformer directement la question en une variable mesurable.

Exemple II.2.4

On s'intéresse à l'incidence du diabète dans la population québécoise. Une manière raisonnable de l'opérationnaliser est de calculer le taux de personnes diagnostiquées diabétiques dans un échantillon représentatif de la population. Ici, le concept "incidence du diabète" est suffisamment précis pour être directement mesuré.

Dans de nombreux cas pratiques, le concept étudié est plus complexe (au sens de composé de plusieurs aspects). Dans ce cas, il faut décomposer le concept en plusieurs aspects plus concrets, puis définir comment mesurer chacun de ces aspects.

Exemple II.2.5

On s'intéresse à la "réussite scolaire" des étudiants universitaires. Ce concept est complexe et peut inclure plusieurs dimensions, telles que les notes académiques, la ponctualité, l'engagement dans les activités parascolaires et l'opinion des professeurs. Pour opérationnaliser ce concept, on peut définir les aspects suivants :

- **Notes académiques** : Mesurées par la moyenne générale ramenée sur 100.
- **Ponctualité** : Mesurée par le nombre de retards ou absences non justifiées. On décide de pondérer chaque absence non justifiée à un cours comme 2 retards.
- **Engagement parascolaire** : Mesuré par le nombre d'heures consacrées aux activités étudiantes par semestre.
- **Opinion des professeurs** : À chaque bulletin semestriel, on classe les appréciations des professeurs en 3 catégories : "positif", "négatif", "neutre", et on calcule un score d'appréciation $= \frac{\text{Nombre de positifs} - \text{Nombre de négatifs}}{\text{Nombre total d'appréciations}}$.

Chaque aspect est lui-même opérationnalisé en une variable mesurable, permettant une évaluation plus complète de la réussite scolaire que la simple moyenne générale. Notons qu'on a effectué de nombreux choix : le nombre d'aspects retenus, la mesure de chacun, l'importance relative consacrée aux absences par rapport aux retards, etc.

§2.1.2

Comme on le voit dans l'exemple précédent, lorsqu'on opérationnalise un concept, il est souvent utile de le décomposer en **dimensions** (aspects ou facettes du concept) et d'associer à chaque dimension un ou plusieurs **indicateurs** (mesures concrètes). Dans l'exemple précédent, on a 4 dimensions ayant chacun 1 indicateur et le résultat final est un ensemble de 4 variables mesurables. Cependant, dans certains cas, on peut avoir envie de combiner plusieurs indicateurs en un indice composite, par exemple pour pouvoir comparer globalement les données entre elles.

Exemple II.2.6

En 1990, le Programme des Nations Unies pour le Développement (PNUD) a décidé d'opérationnaliser le concept de développement humain. Le choix a été fait de décomposer ce concept en trois dimensions : santé, éducation et revenu. Pour chaque dimension, le PNUD a choisi des indicateurs spécifiques :

- Santé : 1 indicateur (à partir de l'espérance de vie à la naissance $\mathbf{e}$),
- Éducation : 2 indicateurs (durée de scolarisation moyenne $\mathbf{d_m}$, durée de scolarisation attendue $\mathbf{d_a}$),
- Revenu : 1 indicateur (revenu national brut par habitant à parité de pouvoir d'achat en dollars américains de 2017 $\mathbf{r}$).

Ces dimensions mesurent des aspects disparates du développement humain : on les utilise pour calculer un unique indice composite, l'Indice de Développement Humain (IDH). D'abord, chaque indicateur est tronqué entre une valeur minimale et une valeur maximale (par exemple l'espérance de vie est bornée entre 20 et 85 ans et on calcule $\mathbf{e}' = \max(20, \min(\mathbf{e}, 85))$). Les valeurs minimales et maximales pour chaque indicateur sont : $0 \leq \mathbf{d_m} \leq 15$ (années), $0 \leq \mathbf{d_a} \leq 18$ (années), $100 \leq \mathbf{r} \leq 75000$ (USD). Ensuite, chaque indicateur est normalisé entre 0 et 1 selon la formule :

$$I_{dimension} = \frac{valeur - valeur_{min}}{valeur_{max} - valeur_{min}}$$

Pour l'éducation, on fait la moyenne des deux indicateurs normalisés pour obtenir un seul indicateur d'éducation. Enfin, l'indice de développement humain (IDH) est calculé comme la moyenne géométrique des trois indicateurs normalisés :

$$\text{IDH} = (I_{sante} \times I_{education} \times I_{revenu})^{1/3}$$

Cet indice composite permet de comparer le niveau de développement humain entre différents pays de manière standardisée.

Méthode II.2.7 : Opérationnaliser un concept

En résumé, pour opérationnaliser un concept, on suit généralement les étapes suivantes :

1. Identifier le concept à étudier.
2. Définir précisément ce que le concept signifie dans le contexte de l'étude et (si nécessaire) le décomposer en dimensions ou aspects plus concrets.
3. Choisir des indicateurs mesurables pour chaque dimension ou pour le concept global.
4. Spécifier l'échelle de mesure et l'unité de chaque indicateur.

5. (Selon le cas) Définir comment combiner plusieurs indicateurs en un indice composite.

À chaque étape, on prend le temps de justifier les choix faits en termes de validité et de fidélité étant donné les objectifs et les contraintes de l'étude.

À la fin du processus d'opérationnalisation, on doit disposer d'une liste claire des informations que l'on veut mesurer pour chacune des unités statistiques étudiées, ainsi que leurs unités de mesure. Chaque indicateur mesuré sur un des sujets de l'étude forme ce que l'on appelle une *variable*[2]. On a également un plan clair de la façon dont on va combiner (ou pas) ces variables. Il reste à les récolter.

II.2.3 Sélection des outils de collecte


§3.2

Une fois les variables opérationnalisées, il faut concevoir les instruments qui permettront de collecter les données. Dans certains cas, on est chanceux, et tout ou partie des données voulues existe déjà. On parle de **données secondaires**, par opposition aux **données primaires**, qui sont collectées spécifiquement pour l'étude en cours. Dans ce cas, "l'outil de collecte" est le protocole d'extraction de ces données des bases où elles sont conservées. Bien que simple[3] d'un point de vue logistique, l'utilisation de données existantes impose souvent des contraintes sur les variables disponibles et leur qualité. Si deux (ou plus) jeux de données sont combinés, il faut aussi s'assurer de la compatibilité des formats et définitions, des populations étudiées, etc. Il est fréquent qu'à la fois des données primaires et secondaires soient utilisées dans une même étude, par exemple pour compléter des données existantes ou pour les valider, ou, a minima, pour les comparer.

Exemple II.2.8

Imaginons que l'on veuille estimer le nombre de personnes de plus de 30 ans assistant à un concert. On peut utiliser les données de billetterie, qui contiennent le nombre de billets vendus. Cependant, ces données ne contiennent pas l'âge des acheteurs. On décide de créer un outil de collecte : à l'entrée du concert, un enquêteur interroge aléatoirement 100 personnes sur leur âge (une seule question, variable quantitative). On calcule ensuite la proportion de personnes de plus de 30 ans dans cet échantillon, et on l'applique au nombre total de billets vendus pour estimer le nombre total de personnes de plus de 30 ans assistant au concert.

2. On reviendra bientôt en détail sur les types de variables possibles.
3. Et encore, c'est vite dit : la publication, le partage et l'interopérabilité des données n'a pas toujours été si simple qu'aujourd'hui. L'amélioration de cet état de fait est d'ailleurs une des raisons du succès récent des méthodes d'intelligence artificielle.

Lorsque les données n'existent pas, il y a plusieurs cas de figure. Soit les instruments de collecte existent déjà (par exemple, des questionnaires standardisés pour mesurer la dépression) mais n'ont pas encore été appliqués à la population d'intérêt, soit il faut créer de nouveaux outils de collecte. Dans ce dernier cas, il faut créer des outils de collecte adaptés aux variables définies lors de l'opérationnalisation ainsi qu'au type d'unités statistiques auxquelles on s'intéresse [4]

Outils de collecte

Analyse de documents Extraction systématique d'informations à partir de sources écrites, visuelles ou audio. Par exemple, un historien peut analyser des lettres anciennes pour extraire des données sur les relations sociales à une époque donnée. Dans le cadre de l'analyse quantitative, il est important de définir une grille d'analyse claire et structurée pour extraire les données de manière cohérente à partir des documents. Bien que l'analyse de documents soit souvent nécessaire, le volume de données demandé pour une analyse quantitative peut rendre cette tâche fastidieuse si elle est faite à la main, surtout si les documents sont nombreux ou complexes. Il arrive donc que l'analyse de documents soit menée programmatiquement, voire à l'aide de techniques d'intelligence artificielle. Pour s'assurer de données fiables, ces méthodes automatisées limitent souvent les nuances de l'analyse, par exemple en se limitant à compter la fréquence d'apparition de certains mots ou expressions, ou en classant les documents dans des catégories prédéfinies.

Questionnaires Ensemble de questions structurées posées aux sujets de l'étude pour recueillir des données sur leurs opinions, attitudes, comportements ou caractéristiques. Les questionnaires sont un outil de collecte de données très courant en sciences humaines et sociales, car ils permettent de collecter rapidement des données auprès d'un grand nombre de personnes. Ils peuvent être administrés en personne, par téléphone, par courrier ou en ligne. On distingue deux types de questions dans les questionnaires :

- *Questions fermées* (le plus souvent et le plus facile à traiter) :
  - Questions dichotomiques : réponses binaires (oui/non, vrai/faux).
  - Questions à choix multiples : plusieurs options de réponse, dont une ou plusieurs peuvent être sélectionnées.
  - Échelles de Likert : échelles de réponse ordonnées, souvent utilisées pour mesurer les attitudes ou les opinions, allant de "Pas du tout d'accord" à "Tout à fait d'accord".
  - Questions à échelle de fréquence, de qualité, etc.
  - Questions à échelle numérique : les répondants choisissent un nombre sur une échelle (par exemple, de 1 à 10) pour indiquer leur niveau d'accord, de satisfaction, etc.

4. Un archéologue ne va pas créer un questionnaire à poser à des pointes de flèches en silex.

  - Questions à échelle de classement : les répondants classent une liste d'options selon un ordre de préférence ou d'importance.
  - Questions à réponse numérique : les répondants fournissent une valeur numérique (par exemple, leur âge, le nombre de livres lus par mois, etc.).
- *Questions ouvertes* : réponses libres, ensuite codées pour analyse. Bien qu'il soit possible de mener une étude quantitative à partir de questions ouvertes, la variété possible des réponses fait souvent préférer les questions fermées, qui sont plus faciles à analyser quantitativement.

Grilles d'observation Formulaires standardisés remplis par le chercheur pour enregistrer systématiquement des comportements ou événements observés. On peut voir les grilles d'analyse de documents comme un cas particulier de grilles d'observation. Par exemple, un chercheur étudiant le comportement de la foule lors d'un événement sportif peut utiliser une grille d'observation pour enregistrer des données telles que le nombre de personnes présentes, les types de comportements observés (applaudissements, chants, etc.), et les réactions à différents moments du match.

Tests standardisés Instruments validés mesurant des capacités, aptitudes ou traits. Souvent, la construction de ces tests est un projet de recherche en soi, nécessitant des études préliminaires pour établir leur validité et fidélité. Une fois la qualité du test établie, il peut être utilisé dans d'autres études pour mesurer le même concept de manière fiable.

Entretiens structurés : On définit un protocole d'entretien : on a des questions prédéfinies posées de manière uniforme à tous les participants, et éventuellement un arbre de décision pour guider les questions en fonction des réponses.

II.3 Collecte des données

Une fois les outils de collecte définis, il reste à procéder à la collecte des données : appliquer les outils de mesure à chaque unité statistique de l'étude. Dans le cas idéal, on mesure toutes les unités de la population d'intérêt : on parle de **recensement**. Cependant, cela est souvent impossible en pratique, pour des raisons de coût, de temps ou de faisabilité. On utilise alors des techniques d'**échantillonnage** pour sélectionner un sous-ensemble dont la composition, pour les mesures qui nous intéressent au moins, reflètent celle de la population. On dit dans ce cas que l'échantillon est **représentatif**.

II.3.1 Recensement

Dans son sens habituel, un recensement d'une population signifie la collecte des données démographiques et socio-économiques de tous les individus d'une population donnée. Par exemple, le recensement de la population canadienne, effectué tous les 5 ans par Statistique Canada, vise à collecter des informations détaillées sur chaque résident du pays. Les données recueillies comprennent des informations sur l'âge, le sexe, la langue parlée, le niveau d'éducation, l'emploi, le revenu, etc. Ces données sont cruciales pour la planification des politiques publiques, la répartition des ressources et la compréhension des tendances démographiques.

§3.3

Dans le langage des méthodes quantitatives, le mot recensement désigne plus généralement la collecte des données pour *toutes* les unités statistiques d'une population donnée. Par exemple, un archéologue peut décider de recenser tous les artefacts trouvés sur un site donné ou un historien peut décider de recenser tous les documents d'une archive spécifique. Le recensement est la méthode de collecte de données la plus complète, car elle permet d'obtenir des informations sur chaque unité de la population. Si le recensement est possible, il va nécessairement fournir les données les plus précises et complètes pour l'analyse quantitative. Cependant, mener un recensement complet peut être infaisable en pratique, pour diverses raisons. On peut au moins considérer les suivantes :

— **Coût** : Le recensement peut être très coûteux en termes de temps, d'argent et de ressources humaines. Par exemple, le recensement national canadien coûte des centaines de millions de dollars à chaque édition.

— **Temps** : Le recensement peut prendre beaucoup de temps, surtout si la population est grande ou dispersée géographiquement. Par exemple, le recensement national canadien prend plusieurs mois pour être complété.

— **Accessibilité** : Certaines unités de la population peuvent être difficiles à atteindre ou à mesurer. Par exemple, certaines populations marginalisées ou isolées peuvent être difficiles à inclure dans un recensement.

— **Perturbation** : Le processus de recensement peut perturber la population étudiée, surtout si la collecte de données est intrusive ou exigeante. Par exemple, un recensement médical peut nécessiter des examens physiques ou des tests invasifs. Dans le cas d'un recensement archéologique, le processus de fouille peut endommager les artefacts ou les contextes stratigraphiques : on pourrait vouloir par exemple broyer des dents trouvées dans un site néolithique pour détecter la présence de la peste : on évite en général de broyer *toutes* celles qu'on trouve.

II.3.2 Échantillonnage

L'échantillonnage est le processus de sélection d'un sous-ensemble représentatif de la population. Il existe plusieurs méthodes, chacune avec ses avantages et limites. La qualité de

l'échantillonnage influence directement la validité des conclusions tirées de l'analyse quantitative. Toutes choses étant égales par ailleurs, plus un échantillon est grand, plus la précision des estimations sera bonne. Dans le cas idéal, on prend comme échantillon la population entière et les conclusions sont parfaites. Ce n'est pas souvent possible en pratique, pour des raisons de coût, de temps ou de faisabilité et il appartient au chercheur de trouver un équilibre entre les ressources disponibles de son étude et la précision désirée, en fonction de la variabilité de la population étudiée.

Exemple II.3.0

(Exemple fictif) Dans un contexte de difficulté du financement des universités en France, deux instituts de sondage A et B veulent estimer la proportion de la population française favorable à une augmentation des frais d'inscription universitaires.

- L'institut A décide de mener un sondage par téléphone en appelant au hasard 1 000 numéros de téléphone fixes.
- L'institut B décide de mener un sondage en ligne en interrogeant 1 000 personnes passant sur la place de la Sorbonne. [a]

Après analyse, l'institut A rapporte 27% de réponses favorables, tandis que l'institut B rapporte 12% de réponses favorables. Dans les deux cas, les méthodes d'échantillonnage induisent un biais sur les réponses :

- L'institut A n'a pas inclus les personnes sans téléphone fixe (jeunes, personnes à faible revenu, etc.), qui pourraient être plus hostiles à une augmentation des frais d'inscription.
- À l'inverse, les étudiants, qui ont intérêt à garder les frais bas, sont très probablement surreprésentés dans l'échantillon de l'institut B.

Évidemment, dans la réalité, les instituts de sondage font tout leur possible pour minimiser ces biais, mais cet exemple illustre l'importance du choix de la méthode d'échantillonnage.

a. Où se situe la Sorbonne, une université parisienne très connue.

Définition II.3.1 : Échantillonnage

L'**échantillonnage** est l'ensemble des techniques utilisées pour sélectionner un échantillon à partir d'une population, de manière que les caractéristiques de l'échantillon reflètent celles de la population.

II.3.3 Méthodes d'échantillonnage probabilistes

§3.5

Le principe fondamental des méthodes d'échantillonnage probabilistes est que chaque unité de la population a une probabilité connue et non nulle d'être incluse dans l'échantillon. Cela

permet de minimiser les biais de sélection et de garantir une forte probabilité que l'échantillon soit représentatif de la population. De plus, les méthodes probabilistes permettent d'estimer la marge d'erreur et la précision des résultats obtenus à partir de l'échantillon.

En contrepartie de cette précision, la majorité des méthodes d'échantillonnage probabilistes nécessitent une liste complète des unités de la population, appelée **cadre** ou **base d'échantillonnage**. Cette liste permet de sélectionner les unités de manière aléatoire selon la méthode choisie. On suppose que pour chaque unité de la population, on peut récolter les données voulues (joindre la personne, accéder à l'artéfact pour le mesurer, lire le document, etc.).

Aléatoire simple On numérote chaque unité de la population, puis on utilise un générateur de nombres aléatoires pour sélectionner les unités (distinctes) à inclure dans l'échantillon. Chaque unité a la même probabilité d'être sélectionnée.

— *Avantage :* Simple, non biaisé, bonnes propriétés théoriques si l'échantillon est assez grand.

— *Inconvénient :* Nécessite une liste complète de la population, et de pouvoir effectivement joindre n'importe quelle unité.

Exemple II.3.2

Je veux tester que les boîtes de nourriture pour bébés dans mon stock de 838 boîtes sont de bonne qualité. Je sélectionne aléatoirement 10 boîtes parmi celles-ci que je vais examiner (pour simplifier, on suppose qu'elles portent des numéros de série consécutifs qui commencent à partir de 1). Je génère 10 nombres aléatoires entre 1 et 838 (sans répétition) :

$$771, 114, 295, 259, 136, 335, 553, 813, 76, 562.$$

Je vais examiner les boîtes portant ces numéros.

Systématique Étant donné une population de taille N et une taille d'échantillon désirée n, on calcule le pas d'échantillonnage (ou intervalle de sélection) $k = \lfloor \frac{N}{n} \rfloor$ (on prend la partie entière si nécessaire). On choisit un point de départ aléatoire [5] i entre 1 et k, puis on sélectionne chaque k^{e} unité à partir de ce point de départ. Les unités qui constituent l'échantillon sont donc :

$$i, i + k, i + 2k, i + 3k, ..., i + (n - 1)k.$$

— *Avantage :* Plus facile à mettre en œuvre que l'aléatoire simple. De plus, si la liste des unités est ordonnée de manière pertinente (par exemple, par région géographique), cette méthode peut améliorer la représentativité de l'échantillon. Enfin, si on ne connait pas à l'avance la taille de la population, et qu'on décide à l'avance de la proportion d'unités

5. i pour initial

à inclure dans l'échantillon, cette méthode permet de s'adapter dynamiquement à la taille réelle de la population.

— *Inconvénient :* Risque de biais si la liste a une structure périodique, par exemple si on a affaire à des données temporelles qui représentent les jours de la semaine et qu'on choisit un intervalle multiple de 7.

Exemple II.3.3

Le jardin botanique de Montréal veut connaitre le niveau de satisfaction des visiteurs des jardins dans la semaine qui suit l'installation d'une nouvelle exposition florale. On veut un échantillon qui représente au moins 1/100 des visiteurs de la semaine, mais on ne sait pas à l'avance combien de visiteurs il y aura. On décide donc d'utiliser un échantillonnage systématique : on choisit un pas $k = 100$ et un point de départ aléatoire entre 1 et 100, disons 37. On interroge donc le 37^{e}, 137^{e}, 237^{e}, ... visiteurs qui entrent dans le jardin pendant la semaine. Si au total 12 438 visiteurs entrent dans le jardin cette semaine-là, on aura interrogé exactement 125 personnes (le 37^{e} jusqu'au 12 437^{e}).

Stratifié On divise la population en sous-groupes homogènes appelés **strates** (par exemple, par sexe, âge, région géographique, etc.). Ensuite, on effectue un échantillonnage aléatoire simple ou systématique au sein de chaque strate. La taille de l'échantillon dans chaque strate peut être proportionnelle à la taille de la strate dans la population (échantillonnage proportionnel) ou fixe (échantillonnage non proportionnel).

— *Avantage :* Assure une représentation adéquate des sous-groupes importants, améliore la précision des estimations.

— *Inconvénient :* Nécessite une connaissance préalable de la population pour définir les strates.

Exemple II.3.4

Je veux obtenir la moyenne de la taille des étudiants de deuxième session à André-Laurendeau. Je sais que la population totale est de 2000 étudiants, dont 1200 femmes et 800 hommes (nombres *complétement* inventés). Je sais que les femmes sont en moyenne plus petites que les hommes. Par commodité, je décide de prendre mon échantillon dans la classe d'analyse quantitative : en effet, les étudiants ne sont pas groupés par taille dans les classes et ma classe a toutes les chances d'être représentative de l'ensemble des étudiants. Malheureusement pour moi, j'ai autant d'hommes que de femmes dans ma classe : mon échantillon n'est pas représentatif de la population totale et une estimation de la taille moyenne basée sur cet échantillon sera biaisée (trop grande). Pour régler ce problème, je décide de faire un échantillonnage stratifié : je sélectionne aléatoirement 24

femmes et 16 hommes dans ma classe, ce qui correspond à la proportion de femmes et d'hommes dans la population totale. Mon échantillon est maintenant plus représentatif et mon estimation de la taille moyenne sera plus précise.
Cet exemple est une combinaison de deux modes d'échantillonnage : j'ai d'abord fait un échantillonnage par convenance (ma classe), puis un échantillonnage stratifié à l'intérieur de celle-ci.

Par grappes La population est divisée en groupes naturels appelés **grappes** (par exemple, des quartiers, des écoles, des entreprises, les élèves d'une rangée dans une classe, etc.). Ensuite, un certain nombre de grappes sont sélectionnées aléatoirement et toutes les unités à l'intérieur des grappes sélectionnées sont incluses dans l'échantillon.

— *Avantage :* Plus économique et pratique pour les populations dispersées géographiquement.
— *Inconvénient :* Possiblement moins précis que les autres méthodes, car les unités à l'intérieur des grappes peuvent être plus similaires entre elles qu'avec le reste de la population.

Exemple II.3.5

Je veux estimer l'intérêt pour les mathématiques de l'ensemble des élèves d'André-Laurendeau. Plutôt que d'envoyer des questionnaires à des élèves choisis aléatoirement dans l'annuaire du cégep. Je veux un échantillon d'au moins une centaine de personnes, ce qui correspond à environ 4 classes. Je sélectionne donc aléatoirement 4 classes parmi toutes les classes du cégep, puis j'envoie le questionnaire à tous les élèves de ces classes. C'est un échantillonnage par grappes.
En examinant les résultats, je suis surpris de constater que l'intérêt pour les mathématiques est très élevé dans mon échantillon. En y réfléchissant, je réalise que j'ai accidentellement sélectionné 3 classes de sciences naturelles et seulement 1 de sciences humaines. La partie aléatoire de mon échantillonnage (le choix des grappes) porte sur un très petit nombre de grappes (4 classes), et le sujet de mon étude (l'intérêt pour les mathématiques) est fortement lié au type de classe. Mon échantillon n'est donc pas représentatif de l'ensemble des élèves du cégep et mon estimation de l'intérêt pour les mathématiques est biaisée (trop élevée).
Pour régler ce problème, je pourrais choisir plus de grappes, ou stratifier mon choix de grappes en fonction du type de classe (par exemple, choisir 2 classes de sciences naturelles et 2 de sciences humaines), mais le problème fondamental vient du fait que, par construction, les classes sont plus homogènes que l'ensemble des élèves du cégep sur le critère de l'intérêt pour les mathématiques. Une question plus adaptée serait par

exemple le temps de trajet domicile-école : les classes n'ont pas de raison d'être plus homogènes sur ce critère que l'ensemble des élèves du cégep.

Comme vous le voyez, les méthodes d'échantillonnage peuvent être combinées entre elles pour s'adapter aux contraintes pratiques de l'étude, en fonction des forces et faiblesses de chaque méthode.

Exemple II.3.6 : Comparaison des méthodes

Contexte : Étude sur 10 000 étudiants d'une université répartis dans 50 programmes

- **Aléatoire simple :** Tirer au sort 500 étudiants parmi les 10 000
- **Systématique :** Prendre 1 étudiant sur 20 dans la liste alphabétique
- **Stratifié :** Diviser par programme, puis échantillonner proportionnellement (si un programme représente 10% des étudiants, il représentera 10% de l'échantillon)
- **Par grappes :** Sélectionner aléatoirement 10 programmes, enquêter tous les étudiants de ces programmes

II.3.4 Méthodes d'échantillonnage non probabilistes

§3.6

Contrairement aux méthodes probabilistes, les méthodes d'échantillonnage non probabilistes ne reposent pas sur le hasard pour sélectionner les unités de l'échantillon. Elles sont souvent utilisées lorsque la constitution d'une liste exhaustive de la population est impossible ou lorsque des contraintes pratiques l'imposent. Cependant, elles présentent un risque plus élevé de biais et ne permettent généralement pas de généraliser les résultats à l'ensemble de la population.

Échantillonnage par convenance On sélectionne les individus les plus facilement accessibles ou disponibles (ex. : interroger les passants dans la rue, les étudiants présents en classe).

- *Avantage :* Simple, rapide, peu coûteux.
- *Inconvénient :* Risque élevé de biais, faible représentativité. Par exemple, si on veut mesurer la qualité nutritionnelle des repas à emporter, décider d'enregistrer les ventes d'une sandwicherie végan dans un quartier très fréquenté par des sportifs va probablement biaiser les résultats.

Exemple II.3.7

En psychologie expérimentale, il est courant d'utiliser des échantillons par convenance, souvent composés d'étudiants universitaires volontaires. Par exemple, un chercheur souhaitant étudier les effets de la privation de sommeil sur la mémoire peut recruter des étudiants de son université qui sont disponibles et intéressés à participer à l'étude. Bien

que cette méthode soit pratique et économique, elle peut introduire des biais, car les étudiants universitaires peuvent ne pas être représentatifs de la population générale en termes d'âge, de mode de vie ou de santé.

Échantillonnage par choix raisonné (ou jugement) Le chercheur choisit délibérément les unités jugées les plus pertinentes ou représentatives selon son expertise.

— *Avantage :* Permet de cibler des cas spécifiques ou typiques.
— *Inconvénient :* La subjectivité du chercheur a un fort impact, faible généralisabilité.

Exemple II.3.8

On cherche à établir un classement des destinations de vacances les plus accueillantes pour les Canadiens en hiver. On se poste dans un aéroport international canadien pendant la saison hivernale et on interroge des voyageurs. On leur demande d'abord s'ils ont visité plus de deux destinations différentes pour faire du tourisme au cours des cinq dernières années. Si oui, on leur demande de classer ces destinations en fonction de leur expérience d'accueil (hospitalité, services, etc.). On recueille ainsi des avis de voyageurs expérimentés, mais l'échantillon est biaisé vers les personnes qui voyagent fréquemment et qui ont les moyens de le faire.

Échantillonnage boule de neige Utilisé pour des populations difficiles à atteindre, on demande à chaque participant de recommander d'autres personnes à inclure dans l'étude.

— *Avantage :* Utile pour accéder à des groupes cachés ou rares (ex. : minorités, réseaux sociaux).
— *Inconvénient :* Risque de biais d'homogénéité, car les personnes recrutées se ressemblent souvent.

Exemple II.3.9

Un journaliste souhaite évaluer la crédibilité d'une affirmation scientifique sur un sujet tellement spécifique qu'il lui est difficile de juger par lui-même si une personne est compétente pour la discuter. Il commence par interroger un chercheur dans le domaine général de la question et lui demande de lui recommander des spécialistes à contacter, auxquels il demande ensuite la même chose. En suivant ce processus, le journaliste construit progressivement un réseau de contacts d'experts dont il peut recueillir l'opinion générale.

Échantillonnage par quotas On définit à l'avance des quotas à respecter pour certaines caractéristiques (sexe, âge, etc.) afin de refléter la structure de la population, puis on sélectionne

les individus par une des autres méthodes non aléatoires jusqu'à remplir chaque quota. C'est l'équivalent non probabiliste de l'échantillonnage stratifié.

— *Avantage :* Permet de contrôler la composition de l'échantillon sur certains critères.

— *Inconvénient :* Sélection non aléatoire à l'intérieur des quotas, donc biais possible.

Remarque II.3.10

Les méthodes non probabilistes sont parfois inévitables, mais il faut être conscient de leurs limites et éviter de généraliser les résultats à l'ensemble de la population sans précautions.

II.4 Analyse des données

L'analyse elle-même est le sujet principal de ce cours, que l'on explorera en détails dans les chapitres suivants. Cependant, avant de plonger dans les méthodes statistiques, il est important de comprendre les étapes préliminaires cruciales qui garantissent la qualité et la fiabilité des résultats obtenus. Cette section présente un aperçu des principales étapes de l'analyse des données, depuis le nettoyage initial jusqu'à l'interprétation des résultats.

II.4.1 Nettoyage et préparation des données

Avant d'analyser les données, il est essentiel de les nettoyer et de les préparer. Cette étape, souvent sous-estimée, peut représenter jusqu'à 80% du temps de la phase d'analyse. Heureusement pour nous, on supposera dans ce cours que les données sont déjà collectées et disponibles sous une forme exploitable (tableur, base de données, etc.). Cependant, il est rare que les données brutes soient prêtes à l'analyse sans un certain travail de préparation. Le nettoyage des données vise à identifier et corriger les erreurs, gérer les valeurs manquantes, et transformer les données pour les rendre cohérentes et adaptées aux analyses statistiques prévues.

Exemple II.4.0 : Nettoyage de données

Données brutes

ID	Âge	Sexe	Note
1	19	M	85
2	999	F	92
3	20	m	-5
4	21	F	
5	18	H	78

$\xrightarrow{\text{nettoyage}}$

Données nettoyées

ID	Âge	Sexe	Note
1	19	M	85
2	ND	F	92
3	20	M	ND
4	21	F	ND
5	18	M	78

Problèmes identifiés :

- ID 2 : Âge = 999 (probablement code pour “manquant”)
- ID 3 : Note = -5 (impossible)
- ID 4 : Note manquante
- ID 3, 5 : Codage incohérent du sexe (m vs M, H vs M)

On a *choisi*[a] de remplacer les âges et notes manquantes par “ND” (Non Disponible) et de corriger le codage du sexe pour le rendre uniforme. On aurait pu choisir d’imputer les valeurs manquantes par la moyenne des âges et des notes, mais dans ce cas précis, vu le petit nombre de données, on préfère garder l’information de manque.

a. Encore une fois, c’est un choix, et pas la seule possibilité.

Méthode II.4.1 : Nettoyage des données

1. **Vérification de la cohérence** : Détecter les erreurs de saisie, valeurs impossibles
   - Ex. : Âge négatif, note supérieure à 100, date de naissance future
2. **Traitement des valeurs manquantes**
   - Suppression des observations incomplètes (si peu nombreuses)
   - "Correction" : remplacement par la moyenne, médiane ou méthode plus sophistiquée. On utilise souvent l'anglicisme *imputation*.
   - Marquage explicite des valeurs manquantes (ex. : code “ND” pour Non Disponible)
3. **Détection des valeurs aberrantes**
   - Identifier les valeurs extrêmes qui pourraient être des erreurs
   - Décider de les conserver, corriger ou exclure (avec justification)
4. **Codage des variables**
   - Standardiser les formats (ex. : dates, texte)
   - Si on traite les données par ordinateur, il peut être utile de transformer les réponses textuelles en codes numériques.
   - Créer des variables dichotomiques ("dummy") pour faciliter l'analyse. Par exemple pour la variable "Couleur préférée" avec les modalités "Rouge", "Bleu", "Vert", on crée trois variables binaires : "Préférence Rouge" (1 si Rouge, 0 sinon), "Préférence Bleu" (1 si Bleu, 0 sinon), "Préférence Vert" (1 si Vert, 0 sinon).
5. **Transformation des variables**
   - Regrouper des catégories (ex. : âges en tranches d'âge)
   - Créer de nouvelles variables (ex. : calculer l'IMC à partir de la taille et du poids)
   - Normaliser ou standardiser si nécessaire
6. **Documentation**
   - Tenir un journal des modifications apportées
   - Créer un dictionnaire des variables pour garder la trace des codages et transformations choisies.

Toute décision de nettoyage doit être documentée et justifiée. La suppression ou modification de données doit être faite avec précaution et transparence pour maintenir l'intégrité de la recherche.

II.4.2 Traitement statistique des données

On abordera en détail les différentes méthodes statistiques dans les chapitres suivants. De nombreuses techniques existent pour décrire, résumer, et analyser les données et le choix de la méthode dépend des objectifs de l'étude, du type de données et des hypothèses sous-jacentes.

II.4.3 Interprétation des mesures

Bien que certaines constructions statistiques soient intuitives (comme la moyenne ou la médiane), d'autres le sont moins (comme l'écart-type, les intervalles de confiance, les tests d'hypothèses, etc.). Il est crucial de comprendre non seulement comment calculer ces mesures, mais aussi comment les interpréter correctement dans le contexte de l'étude. Au cours des prochains chapitres, nous verrons en détail à quelles questions chaque mesure répond, comment les interpréter et quelles précautions prendre pour éviter les erreurs d'interprétation courantes.

II.5 Communication des résultats

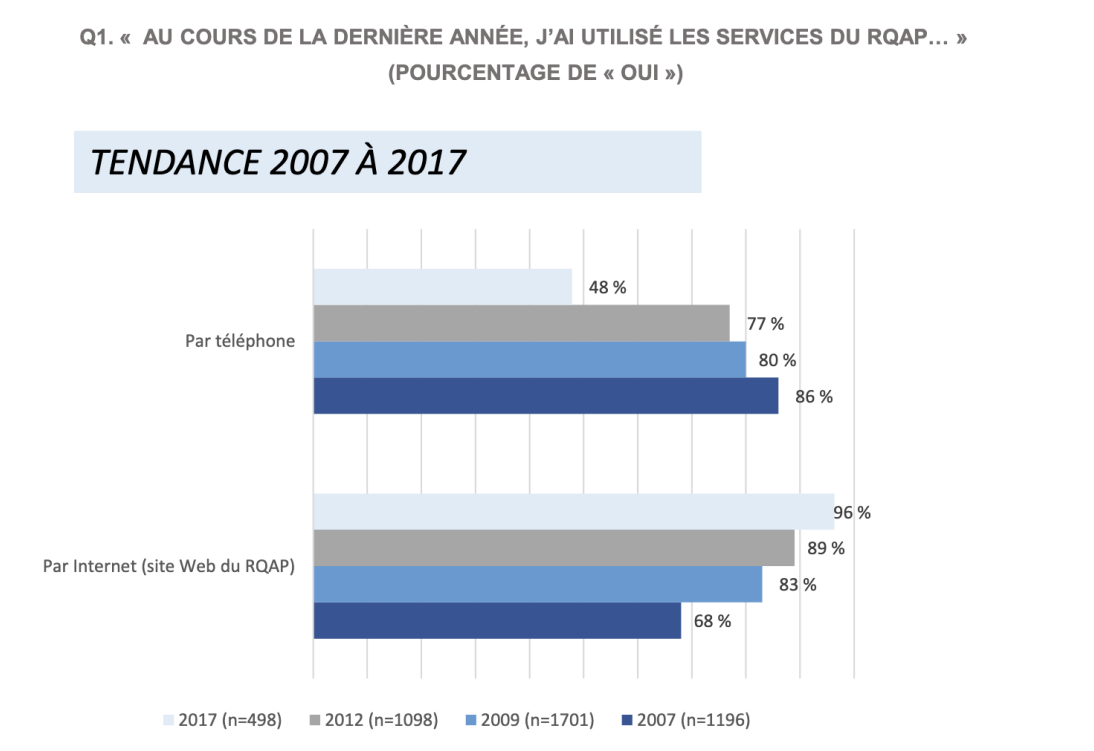
Tout le travail effectué pour collecter, nettoyer, analyser et interpréter les données n'a de valeur que si les résultats sont communiqués de manière claire et efficace. La communication des résultats peut prendre plusieurs formes, en fonction du public-cible et des objectifs de la communication. Les chercheurs tendent à publier leurs résultats dans des articles scientifiques, tandis que les décideurs peuvent préférer des rapports exécutifs ou des présentations visuelles. Les instituts de sondage communiquent souvent leurs résultats via des infographies ou des communiqués de presse.

D'une façon générale, une bonne communication en analyse quantitative donne toutes les informations nécessaires pour que le lecteur puisse comprendre les résultats, évaluer leur validité, et les utiliser pour prendre des décisions éclairées. Cela inclut la description claire de la méthodologie utilisée, la présentation transparente des résultats, et de leur interprétation. En général, cela représente un volume d'information assez important et il est donc important de structurer la communication de manière logique et accessible.

La communication efficace des résultats statistiques eux-mêmes appelle souvent à utiliser des tableaux et graphiques de toutes sortes. Le choix des représentations utilisées dépend du type de données, des résultats à mettre en avant et du type de public visé. Par exemple, si les données sont naturellement ordonnées (comme des dates ou des revenus), on préférera les présenter dans le bon ordre dans une table ou choisir une représentation graphique qui reflète cet ordre. On consacrera un chapitre à l'organisation et à la présentation des données.

Exemple II.5.0

Voici un extrait du *Rapport de sondage sur la satisfaction de la clientèle du régime québécois d'assurance parentale (RAQP)*, publié par le Ministère du Travail, de l'Emploi et de la Solidarité sociale en 2018. [a]



On y trouve :

- La question posée aux répondants au sondage
- Un titre exprimant ce que le graphique illustre.
- Un graphique clair et lisible, avec des couleurs distinctes pour chaque année de réponse, ainsi que les pourcentages exacts, où les données sont regroupées selon la catégorie de réponse.
- Une légende expliquant les couleurs utilisées et précisant la taille de l'échantillon à chaque année.

La méthodologie de collecte et d'analyse étant essentiellement commune à toutes les questions, ces informations sont présentées en début de rapport, qu'on ne reproduit pas ici.

a. Disponible ici : www.rqap.gouv.qc.ca/sites/default/files/documents/publications/RQAP_Rapport_sondage_clientele.pdf

La communication des interprétations des résultats est tout aussi importante. Il faut adap-

ter le langage et le niveau de détail au public cible, en évitant le jargon technique lorsque cela est possible.

Exemple II.5.1

Voici l'interprétation accompagnant le graphique précédent dans le rapport du RAQP :

> *La prestation électronique de service est en croissance*
> *La tendance observée depuis 2007 semble se poursuivre, c'est-à-dire que le mode téléphonique est de plus en plus délaissé au profit du Web. L'utilisation du téléphone pour accéder aux services du RQAP est passée de 86% en 2007 à 48% en 2017, tandis que l'utilisation d'Internet a augmenté de 68% à 96%.*
> *La diminution de l'utilisation du mode téléphonique semble particulièrement importante de 2012 à 2017, puisqu'elle est passée de 77% à 48%. Une explication possible pourrait être que le questionnaire a été administré par Internet en 2017, alors qu'il l'avait été par téléphone lors des sondages précédents. Cela a pu susciter une réponse plus forte chez les utilisateurs réguliers du Web.*

La principale conclusion est clairement énoncée, suivie d'une explication des tendances observées, et d'une réflexion critique sur les possibles biais introduits par la méthodologie de collecte.

Il faut faire attention à ne pas exagérer les conclusions tirées des données. Il est important de distinguer clairement entre corrélation et causalité, et de reconnaître les limites des analyses effectuées. Une communication honnête et transparente renforce la crédibilité des résultats et permet aux lecteurs de les interpréter correctement. Inversement, il ne faut pas non plus "sous-vendre" les résultats : avancés trop prudemment, ils peuvent être ignorés, alors que s'ils sont solides et bien interprétés, ils peuvent fournir des informations précieuses pour la prise de décision.

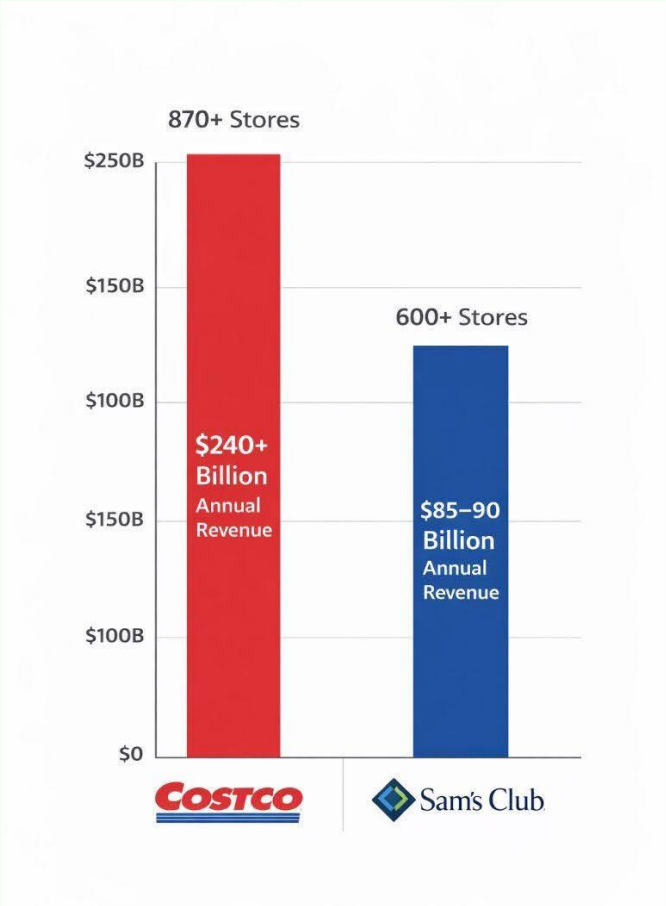
Exemple II.5.2

Le secrétaire à la Santé des États-Unis, Robert Kennedy Junior, a récemment fait retirer du site du centre pour le contrôle des maladies (CDC) les affirmations selon lesquelles les vaccins ne causent pas l'autisme au prétexte qu'aucune étude n'a prouvé que les vaccins ne peuvent pas causer l'autisme. Or, il est extrêmement rare qu'une étude puisse prouver une absence d'effet : tout au plus, une étude peut montrer que l'existence d'un effet est soit très improbable, soit que l'effet est très faible, mais cela ne prouve techniquement pas que l'effet n'existe pas. Cela est dû au fait que la méthode scientifique ne fonctionne par nature que sur ce qui peut être observé : on ne peut pas

prouver l'inexistence de quelque chose, seulement l'absence d'observation de cette chose dans les conditions étudiées.
Dans ces conditions, "l'acceptabilité" des directives du secrétaire Kennedy repose sur l'ignorance du public sur le langage de la méthode scientifique. En dehors du contexte spécifique de la littérature scientifique, il est donc important de communiquer clairement que l'affirmation "les vaccins ne causent pas l'autisme" est aussi certaine qu'il est possible de l'être.

Exemple II.5.3

Pour le plaisir, voici un exemple extrême de mauvaise communication statistique.



Résumé du chapitre

Les étapes de la recherche quantitative

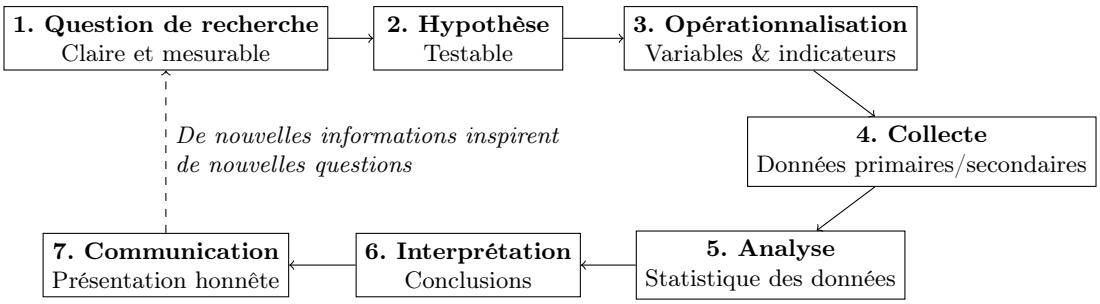


FIGURE II.1 – Processus de recherche quantitative

Opérationnalisation : Transformer le concept en mesure

Processus : Concept théorique → Dimensions → Indicateurs → Variables mesurables
Exemple : "Réussite scolaire"

- **Dimensions :** Notes académiques, ponctualité, engagement, retours des professeurs
- **Indicateurs :** Moyenne générale (0-100), absences non justifiées, heures en activités, score d'appréciation (-1 à +1)

Validité et fidélité : Fondamentaux de la mesure

Validité

- Mesure-t-elle réellement ce qu'on veut mesurer ?
- Plus importante
- Souvent difficile à établir

Fidélité

- La mesure est-elle cohérente et reproductible ?
- Peut exister sans validité
- Plus facile à vérifier

Outils de collecte des données

Outil	Caractéristiques
Données secondaires	Données existantes (bases, registres, archives)
Questionnaires	Questions fermées/ouvertes ; rapide ; large échelle
Entrevues	Flexibilité ; détail ; consommateur de ressources
Observations	Comportements réels ; biais de l'observateur
Analyse de documents	Documents historiques, textes ; codification requise

Chapitre III

Concepts fondamentaux

Avant d'aborder les méthodes statistiques, il est essentiel de maîtriser le vocabulaire et les concepts de base de l'analyse quantitative. Cette section présente les notions fondamentales qui seront utilisées tout au long du cours.

III.1 Le tout vs la partie

Un des buts fondamentaux des statistiques est d'évaluer les caractéristiques d'une population donnée. Malheureusement, dans certains cas [1], il est difficile, coûteux, ou tout simplement

1. En fait, dans la majorité des cas.

impossible de recueillir des données sur l'ensemble de la population. C'est pourquoi on se contente souvent d'étudier un échantillon représentatif de cette population. On se donne donc du vocabulaire permettant de distinguer ce que l'on veut étudier : les *paramètres* de la *population*, de ce que l'on peut réellement mesurer : les *statistiques* de l'*échantillon*.

III.1.1 Population et échantillon

Définition III.1.0 : Population

La **population** (ou univers) est l'ensemble complet de tous les éléments qui font l'objet d'une étude statistique. **Notation** (Taille de la population) : N (pas toujours connu).

Comme le note la définition, la population dépend de l'étude statistique envisagée. Par exemple, si je veux réaliser une étude sur les préférences musicales des élèves du Cégep André-Laurendeau, ma population sera l'ensemble des élèves inscrits dans ce Cégep. L'ensemble de tous les cégépiens du Québec n'est pas ma population pour cette étude spécifique : c'est un ensemble trop grand. De même, l'ensemble des élèves des sections techniques d'André-Laurendeau n'est pas ma population, car je veux inclure tous les élèves, peu importe leur programme d'études.

Exemple III.1.1

Autres exemples de populations :

- Dans le cadre d'une étude sur les intentions de vote à l'élection provinciale, la population est l'ensemble de tous les habitants du Québec (population finie et à un moment donné, fixée).
- Si je veux connaitre le temps moyen d'attente avant d'être servi dans un restaurant, ma population peut être l'ensemble de toutes les commandes passées par des clients dans ce restaurant. Cette population, quoique finie en pratique, est "potentiellement infinie" car de nouvelles commandes peuvent toujours être passées.
- Pour une étude sur le taux de réussite des programmes d'études universitaires au Canada, la population peut être l'ensemble de toutes les filières universitaires offertes dans le pays.

Remarque III.1.2

Attention, comme le montrent les exemples précédents, la population n'est pas toujours un ensemble de personnes. Elle peut aussi être un ensemble d'objets, d'événements, de mesures...

Les populations étudiées dans ce cours seront dans tous les cas finies.

Définition III.1.3

Un **échantillon** est un sous-ensemble de la population, sélectionné pour être étudié.
Notation (Taille de l'échantillon) : n (où $n < N$)

L'échantillon doit être représentatif de la population pour permettre des généralisations valides. Les questions de comment choisir un échantillon représentatif, de mesurer la représentativité, et de corriger les biais d'échantillonnage sont centrales et seront abordées plus tard dans le cours.

Exemple III.1.4

Je veux savoir si les élèves de ma classe de statistiques (population) ont bien compris ce que je viens d'expliquer. Je demande aux élèves du premier rang (échantillon) s'ils ont compris. Cet échantillon n'est probablement pas représentatif de la population, car les élèves du premier rang sont souvent les plus attentifs. A priori, choisir de demander aux élèves dont le nom de famille commence par A, B ou C serait un échantillon plus représentatif, car il n'y a pas de raison que la compréhension des statistiques soit liée à l'ordre alphabétique des noms de famille.

Définition III.1.5 : Unité statistique

L'**unité statistique** (ou individu statistique) est chaque élément de la population ou de l'échantillon sur lequel on effectue des observations.

Exemple III.1.6

Dans une étude sur la performance scolaire, chaque étudiant est une unité statistique.

On a vu dans la partie sur la méthodologie de la recherche scientifique différentes méthodes de sélection d'un échantillon (aléatoire : simple, systématique, stratifié, par grappes et non aléatoire : de convenance, volontaire, boule de neige, par quota). Le choix de la méthode d'échantillonnage dépend de la nature de la population, des ressources disponibles et des objectifs de l'étude.

Définition III.1.7

On appelle **taux d'échantillonnage** (ou taux de sondage) le rapport entre la taille de l'échantillon et la taille de la population, soit $\frac{n}{N}$: c'est la proportion de la population qui est incluse dans l'échantillon. On appelle **poids d'échantillonnage** (ou poids de sondage) l'inverse du taux d'échantillonnage, soit $\frac{N}{n}$: c'est le nombre d'unités de la population que représente chaque unité de l'échantillon.

Exemple III.1.8

Supposons que nous avons une population de 1000 individus et que nous sélectionnons un échantillon de 100 individus. Le taux d'échantillonnage est alors $\frac{100}{1000} = 0.1$ (10%), ce qui signifie que nous avons inclus 10% de la population dans notre échantillon. Le poids d'échantillonnage est l'inverse de ce taux, soit $\frac{1000}{100} = 10$, ce qui signifie que chaque individu de l'échantillon représente 10 individus de la population.

III.1.2 Paramètres et statistiques

Je ne saurais trop insister sur l'importance de distinguer clairement les *paramètres de la population* des *statistiques de l'échantillon*. Encore une fois, une large part du travail du statisticien consiste à estimer les paramètres inconnus de la population à partir des statistiques calculées sur un échantillon.

Définition III.1.9

Un **paramètre** est une valeur numérique qui décrit une caractéristique de la *population*. Les paramètres sont généralement inconnus et notés avec des lettres grecques.

Exemple III.1.10

On verra dans la suite du cours les paramètres suivants (entre autres) :

- μ (mu) : moyenne de la population
- σ (sigma) : écart-type de la population

Définition III.1.11

Une **statistique** est une valeur numérique calculée à partir des données d'un *échantillon*. Les statistiques sont notées avec des lettres latines.

Exemple III.1.12

On verra dans la suite du cours les statistiques suivantes (entre autres) :

- $\bar{x}$: moyenne de l'échantillon
- s : écart-type de l'échantillon

III.2 Variables

Une fois la méthodologie de collecte des données établie et l'échantillon constitué, on commence à récolter les données.

Définition III.2.0 : Donnée et variable

Une **donnée** est une valeur mesurée ou observée pour une variable sur une unité statistique.
Une **variable** est une caractéristique mesurée ou observée sur chaque unité statistique.
Les variables peuvent être classées selon leur nature et leur échelle de mesure.

En d'autres termes, pour chaque unité statistique (individu) de l'échantillon, on mesure ou observe une ou plusieurs caractéristiques (variables) qui prennent des valeurs spécifiques (données).

III.2.1 Variables et types de données

Les variables peuvent prendre différentes natures selon le type de données qu'elles représentent. On distingue principalement deux grandes catégories de variables : les variables qualitatives (ou catégorielles) et les variables quantitatives (ou numériques). Dans chaque catégorie, il existe des sous-types de variables.

Variables qualitatives

Une variable qualitative décrit une caractéristique non numérique qui peut être classée en catégories ou groupes. Les variables qualitatives peuvent être divisées en deux sous-types : nominales et ordinales.

Définition III.2.1 : Variable nominale

Variable dont les catégories n'ont pas d'ordre naturel. Les valeurs sont des étiquettes ou des noms. On peut seulement dire si deux valeurs sont égales ou différentes. On les appelle aussi *variables catégorielles*.

Exemple III.2.2

- Sexe : masculin, féminin, autre
- Province de résidence : Québec, Ontario, Colombie-Britannique, etc.
- Couleur préférée : rouge, bleu, vert, etc.
- État civil : célibataire, marié, divorcé, veuf

Définition III.2.3 : Variable ordinale

Variable dont les catégories ont un ordre naturel, mais sans distance mesurable entre elles. En plus de pouvoir dire si deux valeurs sont égales ou différentes, on peut aussi dire si une valeur est supérieure ou inférieure à une autre.

Exemple III.2.4

- Niveau de scolarité : primaire, secondaire, collégial, universitaire
- Niveau de satisfaction : très insatisfait, insatisfait, neutre, satisfait, très satisfait
- Classement : premier, deuxième, troisième
- Taille de vêtement : XS, S, M, L, XL

Remarque III.2.5

On appelle les valeurs prises par des variables qualitatives des **modalités**.

Variables quantitatives

Une variable quantitative représente une caractéristique mesurable qui peut être exprimée numériquement. Les variables quantitatives peuvent être divisées en fonction des opérations mathématiques qui peuvent être effectuées sur elles (intervalle vs rapport) et de la nature des valeurs qu'elles peuvent prendre (discrète vs continue).

Définition III.2.6 : Variable d'intervalle

Une variable d'**intervalle** est une variable quantitative pour laquelle les différences entre les valeurs sont significatives et mesurables, mais qui ne possède pas de zéro absolu (le zéro est arbitraire). En plus de pouvoir tester l'égalité et l'ordre, on peut mesurer les distances entre les valeurs (au sens de calculer leur différence), mais pas les rapports.

Exemple III.2.7

- Température en degrés Celsius ou Fahrenheit : la différence entre 20 °C et 30 °C est la même qu'entre 30 °C et 40 °C, mais 0 °C ne signifie pas "absence de température". Pour cette raison, on ne peut pas dire que 40 °C est deux fois plus chaud que 20 °C, car en degrés Fahrenheit, ces deux températures correspondent à 104 °F et 68 °F respectivement, ce qui donne un rapport d'environ 1,53. Étant donné que la réalité physique de la température n'a pas changé, si le rapport dépend de l'unité de mesure cela signifie qu'on a une échelle d'intervalle, pas de rapport.
- Dates du calendrier : l'écart entre 2000 et 2010 est de 10 ans, mais l'année 0 n'est pas un point d'origine absolu, c'est un choix arbitraire [a]. Par exemple, dans le calendrier musulman les années ont la même longueur, mais nous sommes en 1447 AH (après l'Hégire).

a. Et d'ailleurs, l'année 0 n'existe pas, on passe directement de -1 à 1.

Définition III.2.8 : Variable de rapport

Une variable de **rapport** (ou ratio) est une variable quantitative qui possède toutes les propriétés de l'échelle d'intervalle, mais avec en plus un zéro absolu qui signifie l'absence totale de la caractéristique mesurée. On peut donc effectuer toutes les opérations arithmétiques, y compris les rapports ainsi que tester l'égalité et l'ordre.

Attention, par "zéro absolu", on entend que le zéro représente l'absence totale de la quantité mesurée ce qui place le zéro à un point fixe et non arbitraire sur l'échelle de mesure. Cependant, on peut tout à fait avoir des valeurs négatives dans une variable de rapport. Par exemple, si on considère le patrimoine total d'une personne (actifs moins passifs), une valeur de -5000 $ signifie que la personne a 5000 $ de dettes nettes, donc une absence de richesse. Le zéro représente bien l'absence de richesse, mais on peut tout à fait être en dessous de ce zéro.

Exemple III.2.9

Taille, poids, âge, revenu, distance parcourue depuis un point fixé, durée depuis un instant fixé, nombre d'objets, score à un examen, etc.

Attention à nouveau : on donne comme exemple pour les variables d'intervalle la date du calendrier et comme exemple pour les variables de rapport le temps écoulé depuis un instant fixé. Cela n'est pas contradictoire : il n'y pas de sens à dire "l'année 2024 est deux fois plus tard que l'année 1012", mais il y a un sens à dire "Alice a attendu 30 minutes depuis son arrivée, deux fois plus longtemps que les 15 minutes d'attente de Bob". Retenez la différence suivante.

Remarque III.2.10

L'échelle d'intervalle a un zéro arbitraire (ex. : 0 °C ne signifie pas absence de température), tandis que l'échelle de rapport a un zéro absolu (ex. : 0 kg signifie absence de masse).

Le tableau III.1 récapitule les différentes échelles de mesure et leurs propriétés.

TABLE III.1 – Échelles de mesure et leurs propriétés

Échelle	**Type**	**Propriétés**	**Opérations**
Nominale	Qualitative	Classification, égalité	$=$
Ordinale	Qualitative	Les précédentes + ordre	$=, <, >$
Intervalle	Quantitative	Les précédentes + distance	$=, <, >, +, -$
Rapport	Quantitative	Les précédentes + zéro absolu	$=, <, >, +, -, \times, \div$

Variables discrètes et continues

En plus d'être d'intervalle ou de rapport, les variables quantitatives peuvent être classées en fonction de la nature des valeurs qu'elles peuvent prendre.

Définition III.2.11 : Variable discrète

Variable qui ne peut prendre qu'un nombre fini ou dénombrable de valeurs, généralement des nombres entiers.

Exemple III.2.12

- Nombre d'enfants dans une famille : 0, 1, 2, 3, etc.
- Nombre d'étudiants dans une classe : 15, 20, 25, etc.
- Nombre de tentatives avant une réussite : 1, 2, 3, etc.
- Score à un examen (sur 100) : 0, 1, 2, ..., 100

Définition III.2.13 : Variable continue

Variable qui peut prendre n'importe quelle valeur dans un intervalle donné.

Exemple III.2.14

- Taille en centimètres : 165,5 cm, 170,2 cm, etc.
- Poids en kilogrammes : 65,3 kg, 72,8 kg, etc.
- Temps de réaction en secondes : 0,25 s, 0,31 s, etc.
- Température en degrés Celsius : 22,5 °C, 23,1 °C, etc.

Certaines variables quantitatives peuvent être techniquement discrètes, mais sont souvent traitées comme continues en raison de la finesse des mesures. Par exemple, le revenu annuel peut être mesuré en dollars (discret), mais il est souvent traité comme une variable continue, car le nombre de valeurs possibles est très élevé, que les différences entre deux valeurs successives sont très petites par rapport à l'échelle globale et les valeurs possibles sont régulièrement espacées. Toutes les combinaisons des types de variables quantitatives sont possibles, comme l'illustre le tableau III.2.

TABLE III.2 – Exemples de variables quantitatives selon le type et l'échelle.

	Discrète	**Continue**
Intervalle	Score à un test de QI (valeurs entières, pas de zéro absolu)	Température en °C (peut prendre toutes valeurs, zéro arbitraire)
Rapport	Nombre d'enfants dans une famille(0, 1, 2, ...)	Taille, poids, durée, revenu (valeurs continues, zéro absolu)

III.2.2 Variables dépendantes et indépendantes

Dans une étude de relation entre variables, on distingue :

Définition III.2.15 : Variable indépendante

La **variable indépendante** (ou variable explicative, prédictive) est la variable que le chercheur manipule ou observe pour en étudier l'effet. Elle est considérée comme la cause dans une relation de causalité.
Notation : Généralement notée X

Exemple III.2.16

- Nombre d'heures d'étude (pour prédire la performance)
- Dose d'un médicament (pour étudier son effet)
- Niveau d'éducation (pour expliquer le revenu)

Définition III.2.17 : Variable dépendante

La **variable dépendante** (ou variable à expliquer, variable réponse) est la variable dont on cherche à expliquer ou prédire les variations. Elle est considérée comme l'effet dans une relation de causalité.
Notation : Généralement notée Y

Exemple III.2.18

- Note à l'examen (dépend des heures d'étude)
- Amélioration de la santé (dépend de la dose du médicament)
- Revenu annuel (dépend du niveau d'éducation)

Il peut y avoir plusieurs variables indépendantes et dépendantes dans une même étude : par exemple, le prix d'une maison (variable dépendante) peut être influencé par plusieurs variables indépendantes telles que la superficie, le nombre de chambres, l'emplacement, etc.

En fonction du nombre de variables en jeu, on distingue :

- **Relation bivariée** : Analyse d'une variable dépendante par d'une variable indépendante.
- **Relation multivariée** : Analyse d'une variable dépendante par deux ou plusieurs variables indépendantes.

Dans ce cours, on se concentrera principalement sur les relations univariées, mais la majorité des concepts peuvent être étendus aux relations multivariées.

Attention : malgré leur nom, il est possible dans une analyse d'une variable dépendante Y par une variable indépendante X, que la variable Y ne "dépende" pas de X au sens habituel

du mot. Dans le cas le plus simple, il se peut qu'on ne trouve aucun lien : par exemple, entre la taille d'un étudiant et sa note finale à ce cours, et Y ne dépend pas de X. Mais supposons par exemple qu'on étudie le lien entre le nombre de visites annuelles à l'opéra (variable dépendante) et la surface de la résidence principale (variable indépendante). Il est tout à fait possible que l'on trouve un lien entre les deux : plus la maison est grande, plus le nombre de visites à l'opéra est élevé. Cependant, cela ne signifie pas que la taille de la maison cause une augmentation du nombre de visites à l'opéra. Il peut y avoir une variable cachée (par exemple, le revenu) qui influence à la fois la taille de la maison et le nombre de visites à l'opéra. Ainsi, même si Y dépend de X dans le modèle statistique, cela ne signifie pas nécessairement une relation causale directe entre les deux variables. On résumera cela par la phrase suivante :

Corrélation n'implique pas causalité.

On reviendra en détail sur les types de liens entre variables dans la partie de statistiques bivariées (tests d'indépendance, corrélation et régression) du cours.

III.3 Organisation des données

III.3.1 Données brutes

Les données brutes peuvent théoriquement prendre toutes sortes de forme : notes manuscrites, enregistrements audio, vidéos, images, etc. Si par exemple l'instrument de mesure est une série d'entretiens guidés avec chacun des membres de l'échantillon, les données brutes peuvent être les notes prises par le chercheur qui mène l'entretien, voire l'ensemble des enregistrements vidéo de l'entretien. Cependant, comme on l'a déjà dit, on suppose dans ce cours que l'on a accès à des données quantitatives déjà "nettoyées" et prêtes à l'emploi, au sens où tous ces formats riches tels que les vidéos, images, enregistrements audio, etc. ont été transformés et encodés de sorte à pouvoir être traités numériquement. Dans ce cas, les données sont généralement présentées sous forme de tableaux que l'on appelle *matrices de données*. On parle aussi souvent de *jeu de données* (dataset en anglais) pour désigner une matrice de données.

Définition III.3.0 : Matrice de données

Une **matrice de données** est une présentation des données sous forme de tableau où chaque ligne représente une unité statistique (individu) et chaque colonne représente une variable mesurée ou observée. On appelle l'ensemble des données stockées dans une colonne une **série de données**, et l'ensemble des données dans une ligne une **observation** ou un **enregistrement**.

Exemple III.3.1

Supposons que l'on fasse une enquête sur le nombre de cafés bus par jour (0, 1, 2, 3, 4+) sur un échantillon de 200 personnes actives et qu'on enregistre aussi leur genre (M/F) et leur statut d'activité (étudiant, employé). La matrice de données pourrait ressembler à ceci :

Individu	Genre	Statut	Cafés par jour
1	M	Étudiant	2
2	F	Employé	1
3	F	Employé	3
4	M	Étudiant	0
...	...	...	...
200	F	Employé	1

Chaque ligne correspond à une personne différente (unité statistique) et chaque colonne correspond à une variable mesurée (genre, statut, nombre de cafés bus par jour), chaque cellule contenant la donnée spécifique pour cette unité et cette variable.

Dans ce cas, qu'appelle-t-on une *base de données* ? En fait, une base de données est un ensemble organisé de données, souvent stockées électroniquement dans un système informatique. Une base de données peut contenir plusieurs matrices de données (ou tables) reliées entre elles. Par exemple, dans une base de données d'une université, on pourrait avoir une table pour les étudiants, une autre pour les cours, et une troisième pour les inscriptions aux cours, toutes reliées entre elles. De plus, la base de données peut être stockée sous une forme plus abstraite ou plus riche que de simples matrices, avec des relations complexes entre les différentes tables. Dans ce cours, on se concentre sur l'analyse de matrices de données individuelles.

III.3.2 Données traitées : distributions de fréquences et tableaux

La matrice de données est la matière première de l'analyse statistique. Cependant, pour analyser efficacement les données, il est souvent nécessaire de les organiser et de les résumer de manière plus structurée, en présentant les données après un premier traitement, par exemple en rassemblant les valeurs similaires et en comptant leur fréquence d'apparition. On appelle cela une *distribution de fréquences*. Les fréquences peuvent être exprimées de plusieurs manières : en effectifs (nombre d'occurrences), en fréquences relatives (proportions), en pourcentages, etc. Techniquement, la distribution de fréquences est l'ensemble des fréquences des différents cas de figures possibles dans les données, indépendamment de leur présentation. Cependant, on présente souvent les distributions de fréquences sous forme de tableaux ou de graphiques pour faciliter l'interprétation.

Types de fréquences :

— **Effectif** (n_i) : Nombre d'observations dans chaque catégorie. Cette quantité est parfois appelée *fréquence brute* ou *fréquence absolue*. On préfèrera parler d'effectif dans ce cours pour éviter toute confusion.

— **Fréquence relative** ($f_i = n_i/N$ ou n_i/n selon le cas) : Proportion d'observations dans chaque catégorie/modalité.

— **Fréquence cumulée** ou **cumulative** : Somme des fréquences jusqu'à une certaine valeur. Formellement, la i-ième fréquence cumulée est définie comme $F_i = \sum_{j=1}^{i} f_j$. En français, cette formule se lit : "la fréquence cumulée de la i-ième catégorie (F_i) est égale à la somme (Σ) des fréquences (f_j) de toutes les catégories inférieures ou égales à i ($j = 1$ à i)". On évite en général de parler de fréquences cumulées pour des variables non ordonnées. La fréquence cumulée permet de savoir quelle proportion des observations se trouve en dessous d'une certaine valeur.

On utilise souvent une majuscule pour les fréquences cumulées (F_i) pour les différencier des fréquences simples (f_i), mais ce n'est pas une règle absolue.

— **Pourcentage** : Fréquence relative × 100. On peut aussi donner la fréquence cumulée en pourcentage.

Exemple III.3.2

Dans notre exemple précédent, bien qu'on ait 200 lignes dans la matrice de données, on a essentiellement seulement $2 \times 2 \times 5 = 20$ combinaisons possibles de genre, statut d'activité et nombre de cafés bus par jour.

TABLE III.3 – Nombre de cafés bus en fonction du genre et de l'activité.

Genre & Statut		**0**	**1**	**2**	**3**	**4+**	**Total**
M	Étudiant	10	10	17	5	2	44
	Employé	8	12	15	10	5	50
F	Étudiant	12	13	19	6	3	53
	Employé	9	14	18	8	4	53
Total		39	49	69	29	14	200

On peut aussi présenter les mêmes données en termes de fréquences relatives ou de pourcentages comme c'est le cas dans le tableau suivant, où l'on omet la distinction entre genre et statut d'activité pour ne se concentrer que sur le nombre de cafés bus par jour.

TABLE III.4 – Nombre de cafés bus par jour

Cafés par jour	Effectif (n_i)	Fréquence relative	Fréquence cumulée	Pourcentage	Pourcentage cumulé
0	39	0,195	0,195	19,50%	19,50%
1	49	0,245	0,440	24,50%	44,00%
2	69	0,345	0,785	34,50%	78,50%
3	29	0,145	0,930	14,50%	93,00%
4+	14	0,070	1,000	7,00%	100%
Total	**200**	**1,00**		**100%**	

On voit dans cette table que 34,5% des personnes de l'échantillon boivent 2 cafés par jour, et que 78,5% boivent au plus 2 cafés par jour.

Tableau de fréquences à une entrée On appelle tableau de fréquences à une entrée un tableau qui présente la distribution de fréquences d'une seule variable. Chaque ligne du tableau correspond à une modalité (ou, dans le cas des variables quantitatives, à un intervalle de valeurs) de la variable et ces classes sont indiquées dans la première colonne. La (ou les) colonne(s) suivante(s) présente(nt) une ou plusieurs des mesures de fréquences (effectifs, fréquences relatives, pourcentages, etc.) pour chaque modalité.

Il est important, comme dans toute présentation de données, de respecter la structure de données : si les modalités ou les classes sont ordonnées, il faut les présenter dans l'ordre.

Exemple III.3.3

Dans le jeu de données `mtcars` (Motor Trend, 1974), on peut présenter le nombre de véhicules en fonction de la consommation en miles par gallon (mpg) comme suit :

TABLE III.5 – Répartition des 32 véhicules du jeu `mtcars` par consommation en mpg

Consommation (mpg)	Nombre de véhicules	Fréquence des véhicules
10-14.99	5	0.16
15-19.99	13	0.41
20-24.99	8	0.25
25-29.99	2	0.06
30+	4	0.13
Total	32	1.01

Notons que dans ce tableau, les classes de consommation sont ordonnées de la plus basse à la plus haute, respectant ainsi la nature ordonnée de la variable quantitative. Par ailleurs, la somme des fréquences est légèrement supérieure à 1 (1.01) en raison de l'arrondissement des valeurs.

Tableau de fréquences à deux entrées Un tableau de fréquences à deux entrées (ou tableau croisé) présente la distribution conjointe de deux variables. Chaque ligne du tableau correspond à une modalité (ou intervalle) de la première variable, et chaque colonne correspond à une modalité (ou intervalle) de la deuxième variable. Les cellules du tableau contiennent les effectifs ou les fréquences pour chaque combinaison de modalités des deux variables. On appelle également ce type de tableau un *tableau de contingence* lorsqu'on travaille avec des variables qualitatives.

Typiquement, comme à la fois les lignes et les colonnes servent déjà à indiquer les modalités ou classes des deux variables, on ne peut indiquer qu'une seule mesure de fréquence par cellule (effectif, fréquence relative, pourcentage, etc.). Il est fréquent de présenter des totaux en marge (lignes et colonnes) pour résumer les distributions de fréquence de chaque variable indépendamment de l'autre. Ainsi, les cases "centrales" du tableau contiennent la **distribution croisée**, ou **conjointe**, tandis que les totaux en marge donnent les **distributions marginales**.

Exemple III.3.4

Voici un exemple réel basé sur le jeu de données `mtcars` (Motor Trend, 1974). On croise le nombre de cylindres et le type de boîte.

TABLE III.6 – Répartition des 32 véhicules du jeu `mtcars` par nombre de cylindres et type de boîte

Cylindres	**Type de boîte**		**Total**
	Automatique	Manuelle	
4	3	8	11
6	4	3	7
8	12	2	14
Total	19	13	32

On observe par exemple que, dans ce jeu, la majorité des voitures 4 cylindres sont à boîte manuelle (8 sur 11), tandis que la plupart des 8 cylindres sont automatiques (12 sur 14).

Tableau de fréquences à plus de deux entrées La création de tableaux de fréquences à plus de deux entrées est possible, mais leur présentation devient rapidement complexe et difficile à interpréter. Dans la pratique, on préfère souvent analyser les relations entre paires de variables à la fois, en utilisant des tableaux croisés à deux entrées, et en complétant l'analyse avec des techniques statistiques multivariées lorsque nécessaire. Cependant, à titre d'exemple, la table III.3 plus haut illustre un tableau de fréquences à trois entrées.

III.3.3 Données traitées : distributions de fréquences et graphiques

Pour le moment, on se réfère au manuel concernant les graphiques. On y verra les différents types de graphiques qui suivent. §3

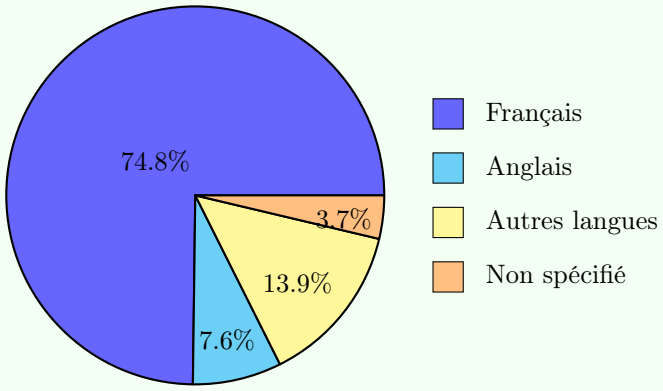
Diagramme en secteurs ou en anneau Constitué d'un disque ou d'un anneau divisé en parts dont la taille est proportionnelle à la fréquence de chaque modalité.

- Adapté pour indiquer les fréquences de variables qualitatives nominales,
- chaque secteur représente une modalité,
- dans le cas des diagrammes en anneau, un cercle vide est présent au centre pour améliorer l'esthétique et la lisibilité, ce qui permet également de comparer plusieurs distributions en imbriquant deux ou plusieurs anneaux.

Exemple III.3.5

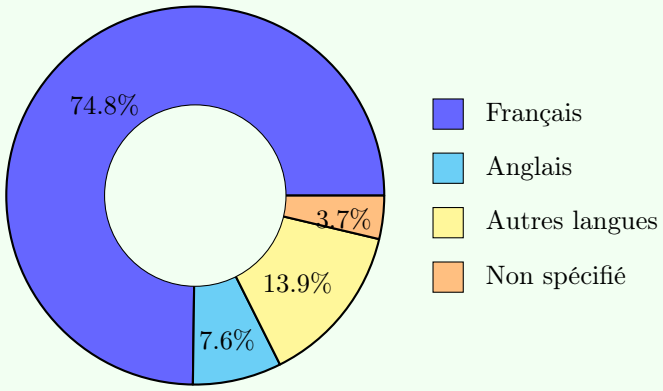
Voici un exemple basé sur des données réelles du recensement canadien de 2021 concernant la répartition des langues maternelles au Québec (en simplifiant les catégories) :

FIGURE III.1 – Répartition des langues maternelles au Québec (2021)



Pour le même ensemble de données, on peut utiliser un diagramme en anneau :

FIGURE III.2 – Répartition des langues maternelles au Québec (2021)



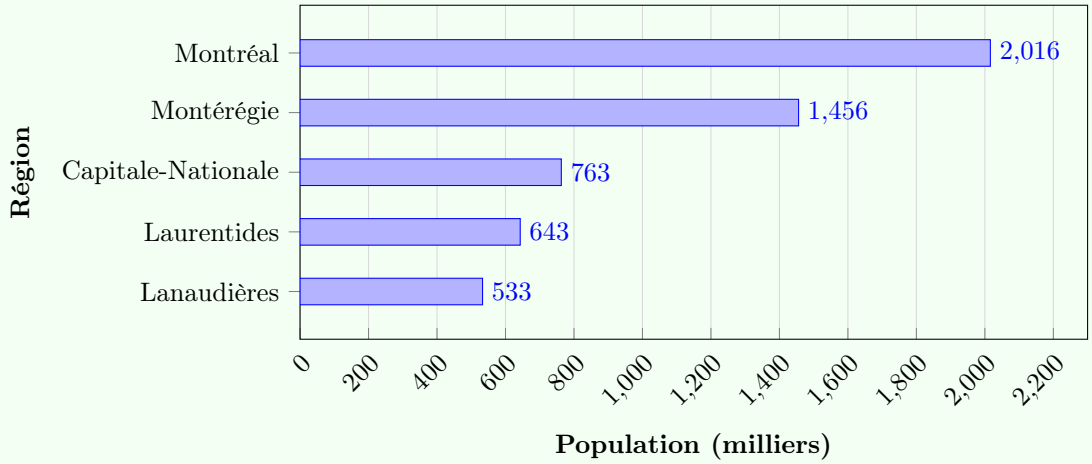
On observe que le français est largement dominant comme langue maternelle au Québec, représentant plus des trois quarts de la population parmi les répondants.

Diagramme en barres (simples, groupées, empilées) Constitué de barres horizontales dont la longueur est proportionnelle à la fréquence de (ou à la quantité associée à) chaque modalité.

— Adapté pour indiquer les fréquences de variables qualitatives nominales et ordinales,
— chaque barre représente une modalité, l'ordre de présentation des barres doit respecter l'ordre naturel des modalités pour les variables ordinales,
— les barres doivent être espacées pour indiquer que les modalités sont distinctes,
— les barres groupées ou empilées sont utilisées pour comparer la distribution d'une variable qualitative en fonction d'une autre variable qualitative.

Exemple III.3.6

FIGURE III.3 – Population des principales régions administratives du Québec (2021, en milliers)

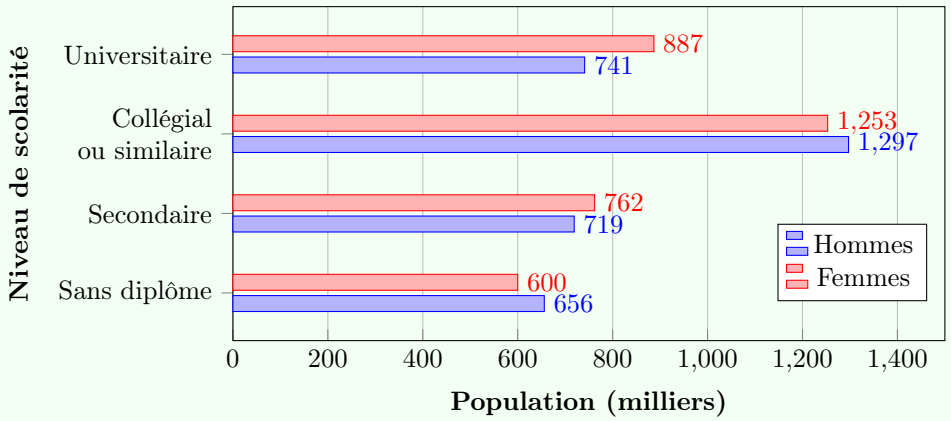


On voit dans ce diagramme en barres que Montréal est de loin la région la plus peuplée, suivie par la Montérégie [a].

a. Source : statistique.quebec.ca/fr/produit/tableau/3595

Exemple III.3.7

FIGURE III.4 – Niveau de scolarité par sexe au Québec (population de plus de 15 ans, 2021, en milliers)

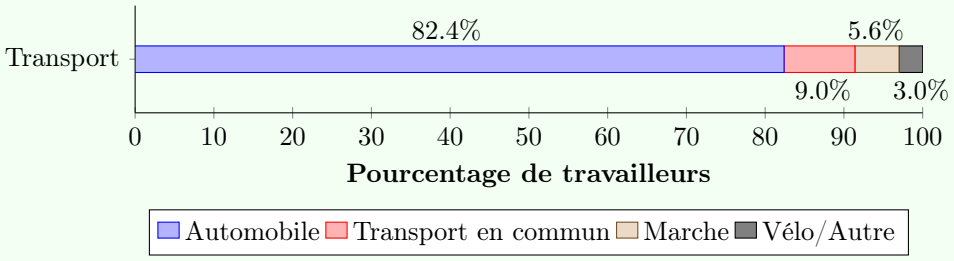


On observe dans ce diagramme en barres groupées que les femmes sont plus nombreuses que les hommes à avoir un diplôme universitaire, tandis que les hommes sont légèrement plus nombreux à s'arrêter au collégial. [a]

a. Source : www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=F

Exemple III.3.8

FIGURE III.5 – Mode de transport pour se rendre au travail au Québec (2021)



On observe dans ce diagramme en barres empilées que l'automobile domine largement comme mode de transport pour se rendre au travail au Québec, représentant environ 80 % des déplacements. [a]

a. Source : www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=F

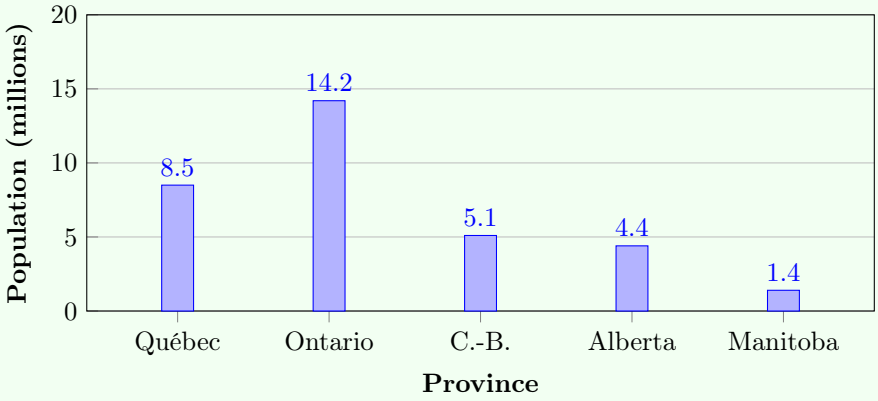
Diagramme en colonnes (simples, groupées, empilées) Constitué de barres verticales dont la hauteur est proportionnelle à la fréquence de (ou à la quantité associée à) chaque modalité.

— Même cas d'usage que les graphiques en barres, mais avec des barres verticales,

— on le préfère aux diagrammes en barres lorsqu'on veut mettre l'accent sur l'ordre des modalités indiquées sur l'axe horizontal.

Exemple III.3.9 : Diagramme en colonnes

FIGURE III.6 – Population de quelques provinces canadiennes (2021, en millions)

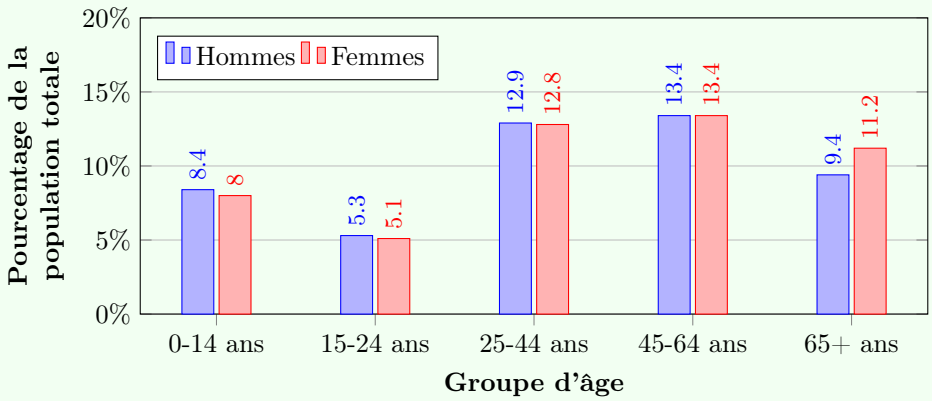


On observe que l'Ontario est la province la plus peuplée, suivie du Québec. [a]

a. Source : www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=F

Exemple III.3.10 : Diagramme en colonnes groupées

FIGURE III.7 – Répartition des Québécois.es par groupe d'âge et par sexe (2021, en %)



On observe que la répartition par groupe d'âge est similaire entre les sexes, avec une

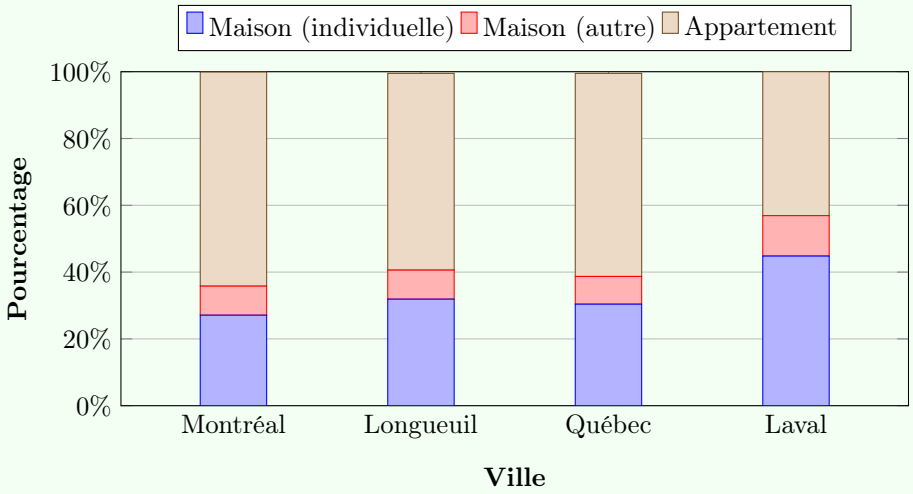
proportion légèrement plus élevée de femmes chez les 65 ans et plus. [a]

a. Source : www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=F

Exemple III.3.11 : Diagramme en colonnes empilées

Voici un exemple de diagramme en colonnes empilées basé sur des données du recensement canadien de 2021 [a] concernant le type de logement par région au Québec (en pourcentage) :

FIGURE III.8 – Composition du parc de logement de quelques villes québécoises (2021)



On observe que la composition du parc de logement à Longueuil est plus semblable à celle de Montréal, que celle de Laval à celle de Québec.

a. Source : www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=F

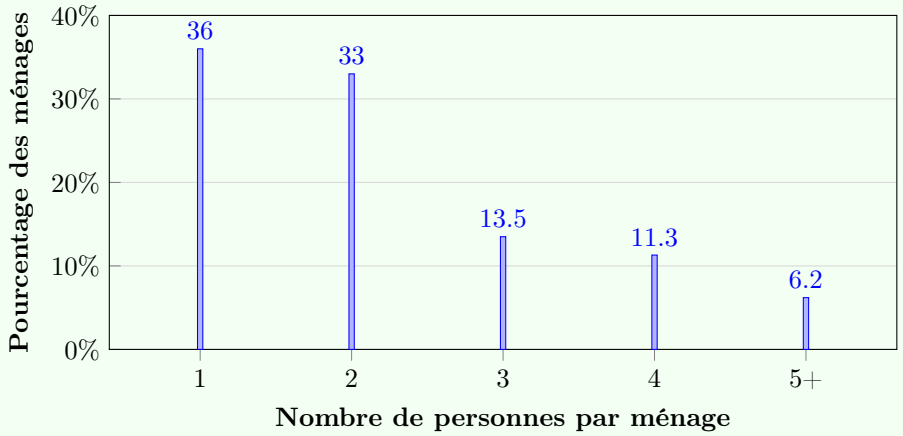
Diagramme à bâtons Constitué de lignes [2] verticales dont la hauteur est proportionnelle à la fréquence de (ou à la quantité associée à) chaque modalité. Ce graphique est très semblable aux graphiques en colonnes, mais il est préféré au diagramme en colonnes pour les variables quantitatives discrètes.

— Chaque bâton représente une valeur possible de la variable,
— les bâtons doivent être espacés pour indiquer que les valeurs sont distinctes,
— utile pour visualiser la distribution des données discrètes et identifier les tendances ou les anomalies.

Exemple III.3.12 : Diagramme à bâtons

Voici un exemple de diagramme à bâtons basé sur des données réelles du recensement canadien de 2021 [a] concernant le nombre de personnes par ménage au Québec (en pourcentage) :

FIGURE III.9 – Distribution du nombre de personnes par ménage au Québec (2021)



On observe que les ménages d'une ou deux personnes sont les plus fréquents au Québec, représentant ensemble environ 69 % des ménages, tandis que les ménages de cinq personnes ou plus sont relativement rares (6,2 %).

a. Source : www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=F

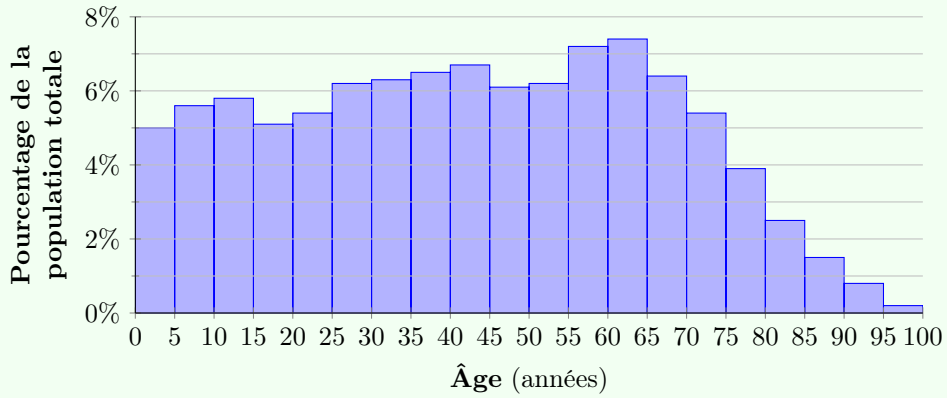
2. Qu'on épaissit souvent en fines colonnes

Histogramme Constitué de barres verticales dont la surface est proportionnelle à la fréquence des valeurs dans chaque intervalle de classe et la largeur de chaque barre est proportionnelle à la largeur de l'intervalle de classe.

— Adapté pour indiquer les fréquences de variables quantitatives continues,
— chaque barre représente un intervalle de valeurs (classe) : la largeur de la barre correspond à l'amplitude de l'intervalle de classe. Dans de nombreux cas, les intervalles de classes sont de largeur égale : dans ce cas, la hauteur de la barre est proportionnelle à la fréquence de l'intervalle de classe. Cependant, lorsque les intervalles de classes ont des largeurs différentes, la hauteur de chaque barre doit être ajustée pour refléter la densité de fréquence (fréquence par unité de largeur) afin que la surface totale de chaque barre soit proportionnelle à la fréquence de l'intervalle de classe.
— Si les classes sont toutes de la même longueur, on peut représenter une autre quantité associée à chaque classe que la fréquence (par exemple, le patrimoine moyen pour des classes de la forme $[20, 25[, [25, 30[, [30, 35[)...$).
— Les barres sont adjacentes pour indiquer que les valeurs sont continues.
— Utile pour visualiser la distribution des données continues, identifier les tendances, la dispersion et la présence de valeurs aberrantes.

Exemple III.3.13 : Histogramme avec des classes de largeur égale

FIGURE III.10 – Distribution de la population du Québec par groupe d'âge (2021, en pourcentage)



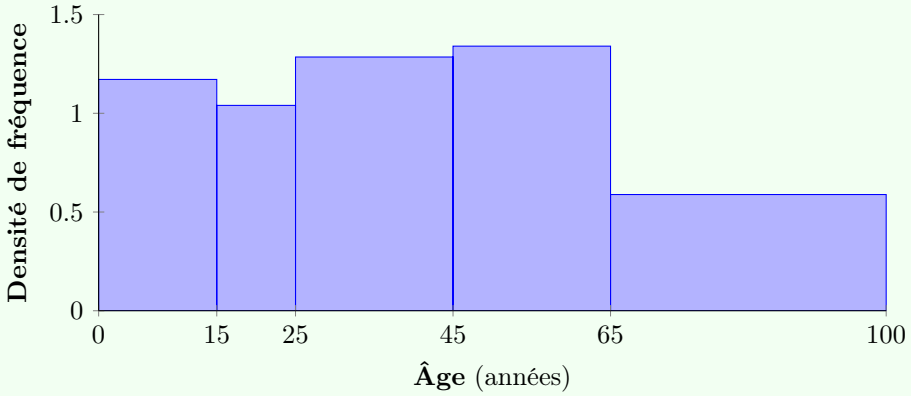
On voit dans ce cas que les groupes d'âge 25-44 ans et 45-64 ans sont les plus nombreux au Québec en 2021. Il y avait 1975 personnes de plus de 100 ans au Québec en 2021 [a], soit environ 0,0 % de la population totale (à une décimale près) et on choisit de ne pas représenter la classe d'âge 100+ dans l'histogramme.

a. Source : www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=F

Exemple III.3.14 : Histogramme avec des classes de largeur inégale

On présente ci-dessous un histogramme de la distribution de la population du Québec par groupe d'âge en 2021 [a], avec des classes d'âge de largeur inégale. L'interprétation de la hauteur des barres est plus difficile : on l'appelle la **densité de fréquence**. Elle représente la concentration des données dans chaque intervalle de classe.

FIGURE III.11 – Distribution de la population du Québec par groupe d'âge (2021, en densité de fréquence)



Par exemple, dans ce cas, la concentration de la population dans les classes 0-65 est à peu près la même (autour de 1,1), tandis que la concentration dans la classe 65+ est plus faible (environ 0,59), reflétant la plus grande dispersion des données dans cette dernière classe que dans les autres. On retrouve que la classe 15-25 est la moins concentrée, ce qui correspond au creux observé dans l'histogramme précédent.

a. Source : www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=F

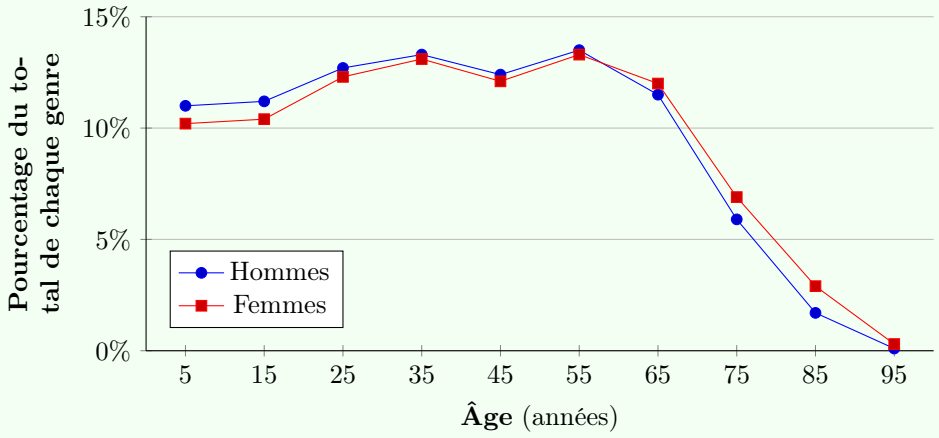
Polygone de fréquence Constitué d'une ligne brisée qui relie les points représentant les fréquences des différentes valeurs ou classes.

— Adapté pour indiquer les fréquences de variables quantitatives discrètes ou continues,
— chaque point représente la fréquence d'une valeur ou d'une classe,
— utile pour visualiser la distribution des données, identifier les tendances et comparer plusieurs distributions.

Exemple III.3.15 : Polygone de fréquence

On présente ci-dessous un polygone de fréquence de la distribution de la population du Québec par groupe d'âge en 2021, avec des classes d'âge de largeur 10 ans.

FIGURE III.12 – Distribution des Québécois.e.s (groupés par genre) par groupe d'âge (2021, tranches de 10 ans, en pourcentage)



On observe que le polygone de fréquence permet de visualiser plus clairement la tendance générale de la distribution, avec un pic autour de 55 ans et une diminution progressive pour les groupes d'âge plus élevés. On voit aussi une inversion de la répartition entre les femmes et les hommes à partir de 65 ans.

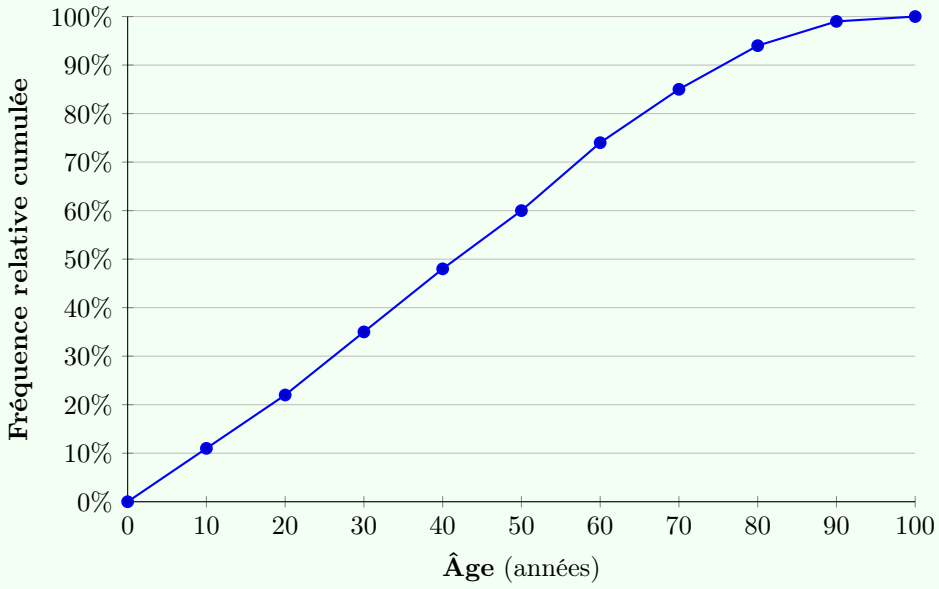
Ogive Constituée d'une courbe qui représente les fréquences (le plus souvent relatives) cumulées des différentes valeurs ou classes.

- C'est la même construction qu'un polygone de fréquence, mais appliquée aux fréquences cumulées, ce qui donne systématiquement une courbe croissante,
- adaptée pour indiquer les fréquences cumulées de variables quantitatives discrètes ou continues,
- chaque point représente la fréquence cumulée jusqu'à une valeur ou une classe,
- si on l'applique aux fréquences relatives cumulées, l'ogive part de 0 et se termine à 1 (ou 0 % à 100 % si on utilise des pourcentages) et permet de lire directement des quantiles (médiane, quartiles, etc.)[3].

Exemple III.3.16 : Ogive

On présente ci-dessous une ogive de la distribution de la population du Québec par groupe d'âge en 2021, montrant les fréquences relatives cumulées.

FIGURE III.13 – Distribution cumulative de la population du Québec par groupe d'âge (2021, fréquences relatives cumulées)



L'ogive permet de lire directement des informations importantes : par exemple, on peut voir qu'environ 50 % de la population québécoise a moins de 42 ans (médiane), que 74 % de la population a moins de 60 ans, et que 94 % a moins de 80 ans. La courbe croissante caractéristique de l'ogive facilite l'identification des quantiles et la comparaison des distributions cumulées : on y reviendra.

3. Que l'on définira plus tard

III.3.4 Un cas particulier : les séries chronologiques

On appelle **série chronologique** une série de données quantitatives collectées au fil du temps, le plus souvent à des intervalles de temps réguliers. Si l'intervalle de temps entre les collectes est constant, on l'appelle **périodicité** de la série. Les séries chronologiques sont utilisées pour analyser les tendances, les cycles et les variations saisonnières dans les données au fil du temps. Les différents types de tableaux et graphiques que l'on a discutés précédemment peuvent être appliqués aux séries chronologiques. Pour les plus adaptés d'entre eux, on utilise souvent un nom différent pour indiquer qu'on travaille avec des données temporelles. Plus généralement, un diagramme indiquant une variable quantitative en fonction du temps est souvent appelé un **chronogramme** ou un **historiogramme**.

Tableau de séries chronologiques : un tableau qui présente les valeurs de la variable mesurée à différents points dans le temps. Chaque ligne du tableau correspond à un point dans le temps (date, heure, etc.) et la (ou les) colonne suivante contient la valeur mesurée à ce moment-là.

Exemple III.3.17

TABLE III.7 – Population estimée de Montréal de 1986 à 2026 (en milliers)

Année	Population (en milliers)	Année	Population (en milliers)	Année	Population (en milliers)
1986	1 820	**2000**	1 833	**2014**	1 948
1987	1 834	**2001**	1 853	**2015**	1 951
1988	1 821	**2002**	1 867	**2016**	1 961
1989	1 836	**2003**	1 871	**2017**	1 984
1990	1 823	**2004**	1 873	**2018**	2 024
1991	1 816	**2005**	1 874	**2019**	2 059
1992	1 799	**2006**	1 877	**2020**	2 062
1993	1 795	**2007**	1 876	**2021**	2 016
1994	1 794	**2008**	1 880	**2022**	2 035
1995	1 795	**2009**	1 895	**2023**	2 095
1996	1 798	**2010**	1 907	**2024**	2 167
1997	1 799	**2011**	1 915	**2025**	2 172
1998	1 801	**2012**	1 927		
1999	1 815	**2013**	1 939		

Notons que dans cet exemple de tableau de séries chronologiques, le nombre d'années

nécessite d'être présenté sur plusieurs colonnes pour éviter que le tableau ne devienne trop large. On peut aussi choisir de présenter les années sur une seule colonne et les populations correspondantes sur une autre colonne, mais cela rendrait le tableau plus long [a].

a. Et plus difficile à mettre en page !

Graphique à lignes brisées : un graphique qui relie les points représentant les valeurs de la variable mesurée à différents points dans le temps avec des lignes droites. Ce type de graphique est particulièrement adapté pour visualiser les tendances et les variations au fil du temps. C'est essentiellement la même construction qu'un polygone de fréquence, mais appliquée aux séries chronologiques sur l'axe horizontal et à une variable quantitative (pas forcément une fréquence) sur l'axe vertical.

Exemple III.3.18 : Graphique à lignes brisées

FIGURE III.14 – Évolution de la population estimée de Montréal (1986–2025, en milliers)

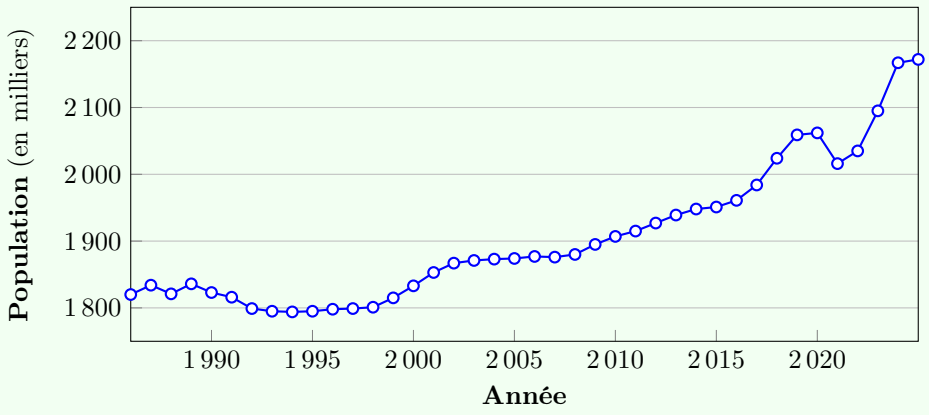


Diagramme en colonnes : similaire à un diagramme en colonnes classique, mais utilisé pour représenter la distribution des valeurs de la variable mesurée sur des intervalles de temps spécifiques. On l'utilise plutôt quand le nombre de périodes est limité et qu'on veut insister sur la comparaison entre ces périodes.

III.3.5 Pourquoi et comment présenter les données ?

L'intérêt de visualiser les données

On l'a vu, plus on récolte un grand nombre de données, plus les informations que l'on va en extraire seront riches et précises. Cependant, à partir d'une certaine quantité de données, il devient impossible d'obtenir une vue d'ensemble en examinant simplement les données brutes. Par exemple, si on a un échantillon de 10 000 individus avec plusieurs variables mesurées pour chacun, il est pratiquement impossible de comprendre la distribution des données ou les relations entre les variables en regardant simplement la matrice de données. C'est là qu'intervient la présentation des données sous forme de tableaux et de graphiques. Ces outils permettent de résumer et de visualiser les données de manière à en extraire des informations clés rapidement et efficacement.

Souvent, d'un seul ensemble de données, on peut extraire plusieurs tableaux ou graphiques, chacun mettant en lumière un aspect différent des données : une variation au cours du temps ou entre groupes, la répartition des mesures en fonction des valeurs, la relation entre plusieurs variables, etc. Chaque tableau ou graphique, dans un certain sens, "élimine" une partie de l'information contenue dans les données brutes pour se concentrer sur un aspect particulier. C'est pourquoi il est important de choisir judicieusement les tableaux et graphiques à utiliser en fonction des questions de recherche que l'on souhaite explorer : un seul graphique peut être trompeur s'il est mal choisi ou mal interprété.

Règles de présentation

Lors de la présentation des données, que ce soit sous forme de tableaux ou de graphiques, il est essentiel de suivre certaines règles pour assurer la clarté et la lisibilité des informations. Voici quelques règles générales à respecter :

- **Titres et légendes** : Chaque tableau ou graphique doit avoir un titre clair et descriptif. Les axes des graphiques doivent être correctement étiquetés avec les unités de mesure le cas échéant. Les légendes doivent expliquer les symboles, couleurs ou styles utilisés.
- **Numérotation** : si votre document comporte plusieurs tableaux ou graphiques, il faut les numéroter (Tableau 1, Figure 1, etc.) pour permettre d'y faire référence sans ambiguïté dans le texte.
- **Ordre logique** : Les modalités ou classes doivent être présentées dans un ordre logique, que ce soit par ordre croissant/décroissant pour les variables ordinales et quantitatives, ou dans un ordre significatif pour les variables nominales.
- **Échelles appropriées** : Choisir des échelles appropriées pour les axes des graphiques afin de représenter fidèlement les données sans distorsion : on évite autant que possible les échelles tronquées ou non proportionnelles qui pourraient induire en erreur. Dans certains cas particuliers, on peut utiliser une échelle logarithmique pour mieux visualiser

des données avec une large gamme de valeurs : il est alors utile d'attirer l'attention du lecteur sur ce choix dans le titre, les étiquettes des axes ou la légende.

— **Simplicité** : Éviter les éléments superflus qui peuvent distraire ou compliquer la lecture des données. La simplicité favorise la compréhension. *On évitera les dégradés de couleurs, les effets 3D, les ombres portées, etc.*

— **Cohérence** : Utiliser des styles, couleurs et formats cohérents à travers tous les tableaux et graphiques pour faciliter la comparaison. Cela permet au lecteur de ne pas avoir à réapprendre la signification des éléments visuels à chaque nouvelle figure.

— **Sources et notes** : Inclure des sources de données et des notes explicatives si nécessaires pour clarifier le contexte ou les méthodes utilisées.

D'une façon générale, le but est de réduire au maximum la charge cognitive du lecteur (et en particulier effacer toute ambiguïté qui pourrait être mal interprétée) et de lui permettre de se concentrer sur le contenu du graphique ou tableau plutôt que de passer du temps à décrypter sa forme. En suivant ces règles, on s'assure que les données sont présentées de manière professionnelle et accessible.

Méthode III.3.19 : Donner un titre approprié

• **À un tableau ou graphique** On pourra suivre la règle générale suivante :

[Ce que montre le graphique ou tableau] sur/des/de/du **[sur quoi ou qui]**
en fonction/selon/par **[variable(s)]** + (année, unités, etc. si nécessaire)

Dans un tableau ou diagramme de fréquences, cela donne plus spécifiquement :

Répartition/Distribution/Composition/Nombre...
des **[unités statistiques]** (**[groupées par X]** si nécessaire)
en fonction de **[variable(s)]** + (année, en pourcentage, en milliers, etc. si nécessaire)

• **À une colonne de tableau à une entrée** Le titre de la première colonne est le nom de la variable considérée, le nom des colonnes suivantes indique le type d'information contenue dans *une* case de la colonne. Dans un tableau de fréquences, cela pourrait être "Nombre/pourcentage/fréquence d'**[unités statistiques]**". Plus généralement, ce pourrait être "Moyenne/Valeur/Maximum/Étendue de **[variable mesurée]**", etc.

• **À une colonne dans un tableau à 2 entrées** Dans ce cas, toutes les lignes et colonnes servent à indiquer une valeur/classe/modalité d'une variable, et le type de données contenu dans les cases est indiqué dans le titre du graphique : par exemple, "Répartition de ... en fonction de ... et de ... (pourcentages)", "Budget moyen de ... en fonction de ... et de ... (en dollars)", etc.

Nom de la variable des lignes	**Nom de la variable des colonnes**			**Total**
	Classe 1	...	Classe k	
Modalité 1	Le type de données des cases est indiqué dans le titre du graphique.			
...				...
Modalité n				
Total		...		

Nombre de classes ou modalités

Un point important à considérer lors de la construction d'un tableau ou d'un graphique est le nombre de modalités (pour les variables qualitatives) ou de classes (pour les variables quantitatives) à utiliser. Si on a un faible nombre de modalités ou classes, on peut se permettre de toutes les inclure dans la représentation. Cependant, si elles sont en nombre trop important elles peuvent rendre la présentation confuse et difficile à interpréter. Dans le cas de variables qualitatives avec de nombreuses modalités, on peut regrouper les modalités moins fréquentes en une catégorie "Autres" pour simplifier la présentation ou choisir une représentation plus adaptée. Par exemple, on considère souvent que les diagrammes en secteurs sont à éviter lorsque le nombre de modalités dépasse 7 et on préfère dans ce cas les diagrammes en barres ou en colonnes.

Pour les variables quantitatives, il est souvent nécessaire de regrouper les valeurs en **classes** (intervalles) pour créer des tableaux ou graphiques lisibles. Soit les données sont déjà regroupées en classes (par exemple, des tranches d'âge), et on peut fusionner certaines classes pour réduire leur nombre si nécessaire. Si, comme c'est plus souvent le cas, il faut créer les classes à partir de données brutes, il faut choisir un nombre de classes approprié. Plusieurs considérations entrent en jeu, mais on choisit le plus souvent des classes de longueur égale, commençant à la valeur minimale observée et finissant à la valeur maximale observée. Le nombre de classes dépend (entre autres) de la taille de l'ensemble de données : plus on a de données, plus on veut de classes, mais on ne veut pas 10 fois plus de classes pour 10 fois plus de données. Une convention répandue est de suivre la **règle de Sturges** : si le nombre total d'observations est n, on choisit le nombre de classes k comme :

$$k = \lceil 1 + 3.22 \log_{10}(n) \rceil$$

où $\lceil x \rceil$ est la fonction plafond qui arrondit x à l'entier supérieur le plus proche. Cela donne le nombre suggéré de classes en fonction du nombre de données suivant :

Une fois le nombre de classes choisi, on choisit un minimum et un maximum (typiquement

TABLE III.8 – Nombre de classes suggéré selon la règle de Sturges en fonction du nombre d'observations

Nombre d'observations (n)	**Nombre de classes suggéré (k)**
Moins de 23	5
De 23 à 45	6
De 46 à 90	7
De 91 à 180	8
De 181 à 361	9
De 362 à 723	10
De 724 à 1447	11
De 1448 à 2895	12
Plus de 2895	13 ou plus

la plus petite et la plus grande valeur observée, arrondies respectivement par défaut et par excès à des valeurs "rondes" pour faciliter la lecture) et on divise l'intervalle entre ces deux valeurs en classes de largeur égale.

Exemple III.3.20

Si les données vont de 12 à 98 et que l'on a 500 observations, la règle de Sturges recommande de choisir 10 classes. On choisit alors des classes de largeur égale de 10 unités, commençant à 10 (arrondi par défaut de 12) et finissant à 100 (arrondi par excès de 98). On calcule une étendue de classe de $90/10 = 9$. Les classes sont alors $[10, 19[, [19, 28[, [28, 37[, [37, 46[, [46, 55[, [55, 64[, [64, 73[, [73, 82[, [82, 91[, [91, 100[$.
Notons qu'on a eu de la chance, car le nombre de classes suggéré divisait proprement l'étendue "arrondie" des données. Considérons des données discrètes allant de 0 à 100 avec 6 classes. L'étendue calculée est alors de $100/6 \approx 16,7$. On peut alors choisir soit d'arrondir l'étendue à 17 et avoir les 6 classes $[0, 17[, [17, 34[, [34, 51[, [51, 68[, [68, 85[, [85, 102[$.

Pour rappel, la notation $[a, b[$ signifie que l'intervalle de classe inclut toutes les valeurs de a à b, incluant a, mais excluant b.

Résumé du chapitre

Distinction population–échantillon

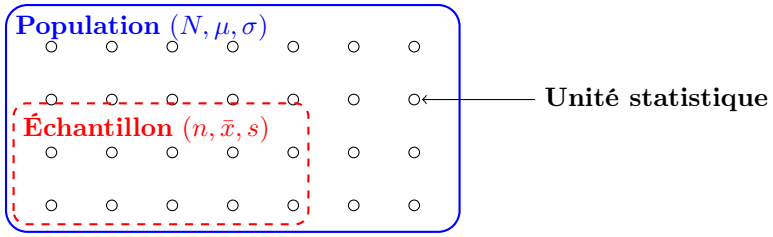


FIGURE III.15 – Relation entre population et échantillon

Vocabulaire essentiel

Terme	Définition
Paramètre	Caractéristique de la **population** (souvent inconnu, noté avec lettres grecques : μ, σ)
Statistique	Caractéristique calculée d'un **échantillon** (souvent mesurable, noté avec lettres latines : $\bar{x}, s$)
Unité statistique	Un individu ou élément dans la population/échantillon
Variable	Caractéristique mesurée sur chaque unité
Donnée	Valeur spécifique d'une variable pour une unité

Échelles de mesure et opérations permises

Échelle	Propriétés	Opérations
Nominale	Classification seulement	$=$ (égal ?)
Ordinale	Ordre naturel	$=, <, >$ (rangement)
Intervalle	Distance mesurable	$=, <, >, +, -$ (différences)
Rapport	Zéro absolu	$=, <, >, +, -, \times, \div$ (rapports)

Variables discrètes vs continues

Discrètes

- Nombre fini de valeurs
- Souvent entières
- Exemples : nombre d'enfants, nombre de voitures

Continues

- Infinité de valeurs possibles
- Nombres décimaux
- Exemples : taille, poids, température

Arbre de décision : Types de variables

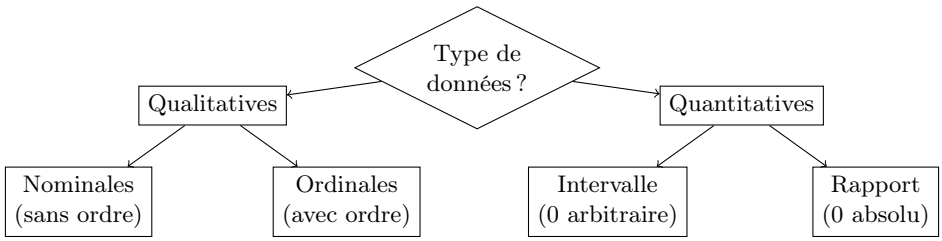


FIGURE III.16 – Classification des types de variables

III.4 Interlude : comparer des grandeurs

III.4.1 Comparer avec des quotients

Définition III.4.0 : Proportion

La **proportion** d'une partie P d'un ensemble E est le rapport $\frac{\text{taille de } P}{\text{taille de } E}$. La proportion est un nombre compris entre 0 et 1, sans unité.

La proportion compare des grandeurs de même nature : des personnes à des personnes, des dollars à des dollars, etc : les unités se simplifient dans le quotient. Comme on compare toujours une partie à un tout, la proportion est toujours comprise entre 0 et 1 : par exemple, la proportion de femmes dans une classe ne peut pas être supérieure à 1 (100%) ni inférieure à 0 (0%).

Exemple III.4.1

Si dans une classe de 30 étudiants, 18 sont des femmes, la proportion de femmes dans la classe est de $\frac{18}{30} = 0,6$.

Définition III.4.2 : Taux

Le **taux** est une proportion exprimée en pour cent, pour mille, pour 10 000, etc.

On utilise des taux exprimés en pour cent/pour mille/pour plus pour rendre les proportions plus lisibles : par exemple, il est plus facile d'interpréter un taux d'échec de 16% à un examen qu'une proportion d'échec de 0,16. Il est plus facile de discuter du taux d'incidence d'une maladie rare comme valant de 3 pour 1 000 000 plutôt que de 0,000003, ou même 0,0003 %. Typiquement, on choisit donc le taux de manière que les chiffres soient compris entre 1 et 100, pour faciliter l'interprétation.

Exemple III.4.3

La prévalence (nombre de cas/taille de la population) de l'autisme au Canada est d'environ 2%, tandis que le taux d'incidence (nombre de nouveaux cas/taille de la population) est d'environ 3‰ (lu 3 pour 1000).

Définition III.4.4 : Ratio

Le **ratio** de deux grandeurs A et B est le rapport $\frac{A}{B}$. On le note aussi $A : B$ (lu "A pour B"). Les unités du ratio sont les unités de A divisées par les unités de B.

Exemple III.4.5

Si dans un groupe de 27 personnes, 9 ont une voiture, le ratio voitures/personnes est de $9 : 27 = \frac{9}{27} = \frac{1}{3}$, soit 1 voiture pour 3 personnes.

Le ratio permet de comparer des grandeurs de nature différente : des voitures par personnes, des dollars par habitant, des professeurs par classe, etc. Si toutefois les grandeurs comparées sont de même nature on peut dans certains cas passer du ratio à la proportion :

$$\text{Un ratio de } A : B \text{ donne une proportion de } \frac{A}{A+B},$$

$$\text{une proportion de } \frac{A}{B} \text{ donne un ratio de } A : (B - A).$$

Exemple III.4.6

Si dans une classe de 30 étudiants, 18 sont des femmes, le ratio femmes/hommes est de $18 : 12 = \frac{18}{12} = \frac{3}{2}$, soit 3 femmes pour 2 hommes, et on a bien que la proportion de femmes est de $\frac{3}{3+2} = 0,6$.

Définition III.4.7 : Indice (élémentaire/synthétique) à base 100

Un **indice à base 100** est un ratio de la forme $\frac{V}{V_0} \times 100$, où V est la valeur de la grandeur que l'on veut comparer, et V_0 est une valeur de référence. L'indice à base 100 est un nombre sans unité. On dit que l'indice est "élémentaire" si V et V_0 sont des valeurs d'une même grandeur à deux points différents (par exemple, le prix d'un panier de biens en 2025 et en 2020), et que l'indice est "synthétique" si V et V_0 sont des valeurs d'une grandeur synthétisant plusieurs grandeurs élémentaires (par exemple, l'indice des prix à la consommation, qui synthétise les prix de plusieurs biens).

III.4.2 Comparer avec des différences

Définition III.4.8 : Variation

La **variation** d'une grandeur V entre deux points A et B est la différence $V(B)-V(A)$. On la note ΔV.

Notez que l'on calcule la valeur de V en B moins celle de V en A : on fait "valeur à l'arrivée" moins "valeur au départ". Par conséquent, si V a augmenté entre A et B, ΔV est positif, et si V a diminué, ΔV est négatif.

La variation d'une grandeur est exprimée dans les mêmes unités que la grandeur elle-même. Par exemple, si V est une population, ΔV est exprimé en nombre d'habitants. Si V est un revenu, ΔV est exprimé en dollars.

Exemple III.4.9

Si entre 2025 et 2026, votre salaire est passé de 50 000 $ à 55 000 $, la variation de votre salaire est de $\Delta V = 55000 - 50000 = 5000$ $. Votre salaire a augmenté de 5 000 $ entre 2025 et 2026.

Cependant, la variation "absolue" peut être difficile à interpréter : par exemple, une personne qui travaille pour la première fois pendant l'été et dont les revenus passent de 0 $ à 5 000 $ a une variation de 5 000 $, tout comme la personne qui gagne déjà 50 000 $ et dont le salaire passe à 55 000 $. Pourtant, la première personne a vu son revenu augmenter de manière beaucoup plus importante que la seconde. Pour résoudre ce problème, on parle de *variation relative*.

Définition III.4.10 : Variation relative

La **variation relative** (ou taux de variation) d'une grandeur V entre deux points A et B est le rapport $\frac{V(B)-V(A)}{V(A)}$. On le note $\frac{\Delta V}{V(A)}$.

Le point de référence, par lequel on divise la variation, est la valeur de V au point de départ A. La variation relative est sans unité, car c'est un rapport de deux grandeurs exprimées dans les mêmes unités. Par exemple, si V est une population, la variation relative est un nombre sans unité, qui peut être interprété comme un pourcentage d'augmentation ou de diminution de la population. On peut l'exprimer en pourcentage, auquel cas, la variation relative est :

$$\frac{V(B) - V(A)}{V(A)} \times 100\%.$$

Exemple III.4.11

Si V a augmenté de 5 000 $ entre 2025 et 2026, et que le salaire initial était de 50 000 $, la variation relative est de $\frac{5000}{50000} = 0,1$, soit une augmentation de 10 %.

Enfin, si votre salaire est passé de 50 000 $ en 2025 à 55 000 $ en 2026, c'est une augmentation très satisfaisante, plus que si votre salaire était passé de 50 000 $ à 55 000 $, de 2020 à 2025 : c'est la même augmentation de 10%, mais sur une période plus courte. Pour quantifier cela, on parle de *variation moyenne*.

Définition III.4.12 : Variation moyenne

La **variation moyenne** d'une grandeur V entre deux temps t_0 et t_1 est le rapport de la variation absolue à la durée de l'intervalle de temps entre t_0 et t_1. On la note $\frac{\Delta V}{\Delta t}$, où Δt est la durée entre t_0 et t_1.

Notez que la variation moyenne est calculée entre deux temps, alors que dans les autres cas, la variation est calculée entre deux points, qui peuvent être des temps, mais aussi d'autres types de points (par exemple, des lieux). L'unité de la variation moyenne est l'unité de V divisée par l'unité de temps. Par exemple, si V est une population exprimée en habitants, et que le temps est exprimé en années, la variation moyenne est exprimée en habitants par an.

Exemple III.4.13

Si votre salaire est passé de 50 000 $ en 2020 à 55 000 $ en 2025, la variation absolue est de 5 000 $, la variation relative est de 10 %, et la variation moyenne est de $\frac{5000}{5} = 1000$ $ par an. La même augmentation de salaire sur 1 an au lieu de 5 ans correspond à une variation moyenne de 5 000 $/an, soit une augmentation beaucoup plus rapide.

Parler d'une variation d'un pourcentage. Imaginons que le taux de chômage est passé de 5 % à 10 %. Comme la grandeur dont on parle est déjà un pourcentage, il pourrait être confus de dire que le taux de chômage a augmenté de 5 % : est-ce que cela signifie que le taux de chômage est passé de 5 % à 10 % (augmentation de 5 points de pourcentage), ou que le taux de chômage a augmenté de 5 % par rapport à son niveau initial de 5 % pour atteindre 5,25 % (augmentation de 0,25 point de pourcentage) ?

Par convention, quand on parle du pourcentage en tant que nombre sans unité, on parle de points de pourcentage. Par conséquent, dans notre exemple, on dira que le taux de chômage a augmenté de 5 points de pourcentage, et non pas de 5 %. Inversement, si on ne précise pas "points de pourcentage", on suppose que l'on parle d'une variation relative, c'est-à-dire d'une augmentation de 5 % par rapport à 5 %, ce qui correspond à une augmentation de 0,25 point de pourcentage.

III.4.3 Indicateurs démographiques

La *variation de population* est la variation absolue de la population entre deux points dans le temps.

$$\Delta P = P(t+1) - P(t),$$

où t est généralement exprimé en années. On peut décomposer la variation de la population en *solde naturel* (naissances - décès) et *solde migratoire* (immigration - émigration) :

$$\Delta P = (N - D) + (I - E),$$

où N est le nombre de naissances, D le nombre de décès, I le nombre d'arrivées (immigration) et E le nombre de départs (émigration) entre les deux points dans le temps.

Exemple III.4.14

Pour 2024, le solde migratoire est de 156700 personnes et le solde naturel de -1400 personnes (il y a eu plus de décès que de naissances).

Si on veut des mesures relatives, on peut considérer le *taux d'accroissement démographique* :

$$\frac{\Delta P}{\text{population initiale}} = \frac{P(t+1) - P(t)}{P(t)} \times 100\%$$

qui se décompose en *taux d'accroissement naturel* :

$$TAN = \frac{N - D}{P(t)} \times 100\%$$

et *taux d'accroissement migratoire* :

$$TAM = \frac{I - E}{P(t)} \times 100\%.$$

Exemple III.4.15

Au Québec, en 2024, selon l'Institut de la statistique du Québec, le taux d'accroissement démographique est de 17,2‰.

Le taux d'accroissement naturel se décompose lui-même en *taux de natalité* $TN = \frac{N}{P(t)} \times 1000\%$ moins le *taux de mortalité* : $TM = \frac{D}{P(t)} \times 1000\%$.

Exemple III.4.16 : L

e taux de natalité au Québec en 2024 était de 8,5‰.

Étant donné que seules les femmes peuvent donner naissance à des enfants, on peut aussi calculer le *taux de fécondité* :

$$TF = \frac{N}{\text{nombre de femmes entre 14 et 49 ans}} \times 1000\%.$$

On peut spécialiser encore davantage le taux de fécondité en calculant le *taux de fécondité par âge* :

$$TFA(a) = \frac{N(a)}{\text{nombre de femmes d'âge } a} \times 1000\%,$$

où $N(a)$ est le nombre de naissances attribuées à des femmes d'âge a.

À partir de ces taux, on peut calculer l' *indice synthétique de fécondité* :

$$ISF = \frac{1}{1000} \sum_{a=14}^{49} TFA(a).$$

Ce nombre représente le nombre moyen d'enfants qu'une femme aurait au cours de sa vie si la naissance de ses enfants était répartie de manière identique à la répartition actuelle des naissances par âge.

Exemple III.4.17

En 2024, l'ISF au Québec était de 1,33, le plus bas jamais enregistré.

Si on imagine que dans une population il y a exactement autant de femmes que d'hommes, et que chaque femme a exactement 1 enfant, chaque personne a $1/2$ descendant à la génération suivante : la taille des générations est divisée par deux à chaque fois, et au fur et à mesure que les générations les plus anciennes disparaissent, la population diminue. Inversement, si chaque femme a 4 enfants, chaque personne a 2 descendants à la génération suivante : la taille des générations est multipliée par deux à chaque fois, et au fur et à mesure que les générations les plus anciennes disparaissent, la population augmente.

Mathématiquement, le point d'équilibre où chaque génération est aussi grande que la précédente correspond à 2 enfants par femme. Dans la réalité, pour tenir compte du fait que tout le monde n'est pas fertile, que certaines personnes meurent avant d'avoir des enfants, etc., on considère que le *taux de remplacement des générations* dans un pays développé comme le Canada est d'environ 2,1 enfants par femme. Par conséquent, si l'ISF est inférieur à 2,1, la population tend à diminuer, et si l'ISF est supérieur à 2,1, la population tend à augmenter.

Deuxième partie

Traitement statistique

Chapitre IV

Statistiques descriptives

Les ensembles de données peuvent comprendre jusqu'à des milliards de données individuelles et si la représentation graphique de celles-ci permet d'en gagner une appréciation qualitative, il est vite nécessaire de trouver un moyen de résumer l'information qu'elles contiennent. C'est le rôle de *statistiques descriptives* : des grandeurs que l'on calcule sur un en ensemble de donnée pour formaliser certaines caractéristiques intuitives : où se trouvent nos observations ? Comment se répartissent-elles ?

IV.1 Mesures de tendance centrale

Les mesures de tendance centrale permettent de résumer un ensemble de données par une valeur représentative. Il y a plusieurs manières, mathématiquement, de définir le "centre" d'une

distribution de données, qui ont toutes leur utilité dans différents contextes. Les trois mesures les plus courantes, que nous allons étudier, sont le mode, la médiane et la moyenne.

IV.1.1 Mode

Définition IV.1.0 : Mode

Le mode est la valeur (ou la modalité, ou la classe) qui apparaît le plus fréquemment dans une distribution.

Exemple IV.1.1

Un magasin de vêtements a vendu au cours d'une journée les tailles suivantes :

S, M, M, L, M, XL, M, S, L, M, M, XL, S, M, L, M, S, M, L, M, L.

Ce qui donne les comptes suivants :

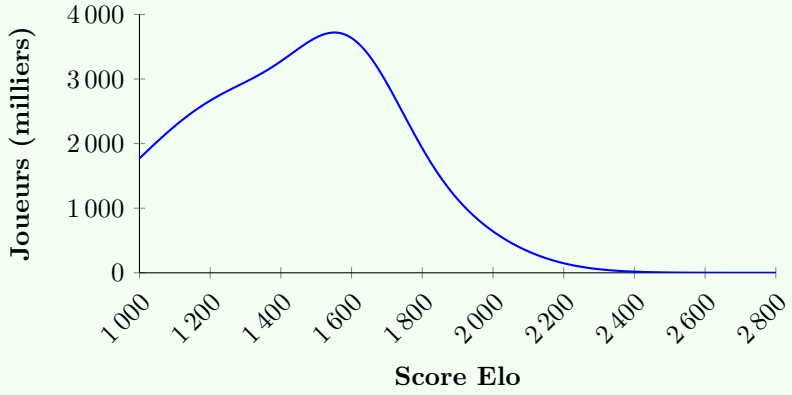
Taille	S	M	L	XL
Nombre vendu	4	9	5	2

Le mode est donc la taille M, qui a été vendue 9 fois, soit plus que toute autre taille.

Attention, le mode n'est pas la *fréquence la plus élevée*, mais bien la *valeur* (ou la classe) associée à cette fréquence. Mathématiquement, le mode est toujours défini, quel que soit le type de données, mais pas forcément unique. Formellement, s'il y a deux modes, ils devraient avoir une fréquence exactement égale. En pratique, les données réelles sont rarement aussi parfaites, et on peut considérer qu'une distribution est *bimodale* si elle présente deux pics distincts, même si les fréquences ne sont pas exactement égales. De la même façon, si le mode ne se distingue pas particulièrement et que toutes (ou au moins une grosse partie) ont des fréquences similaires, on pourra considérer la distribution comme *amodale* : sans mode. On distingue ainsi plusieurs types de distributions selon le nombre de modes, illustrées par les exemples suivants.

Exemple IV.1.2 : Distribution unimodale

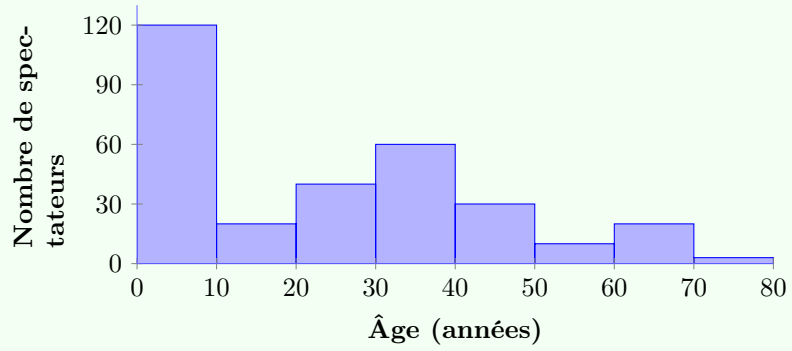
FIGURE IV.1 – Distribution des joueurs d'échecs FIDE en fonction de leur scores Elo.



La figure IV.1 illustre la distribution des scores Elo des joueurs d'échecs enregistrés auprès de la FIDE. On observe un pic prononcé autour de 1600 : cette distribution est *unimodale*, car elle présente un seul mode clairement identifiable.

Exemple IV.1.3 : Distribution bimodale

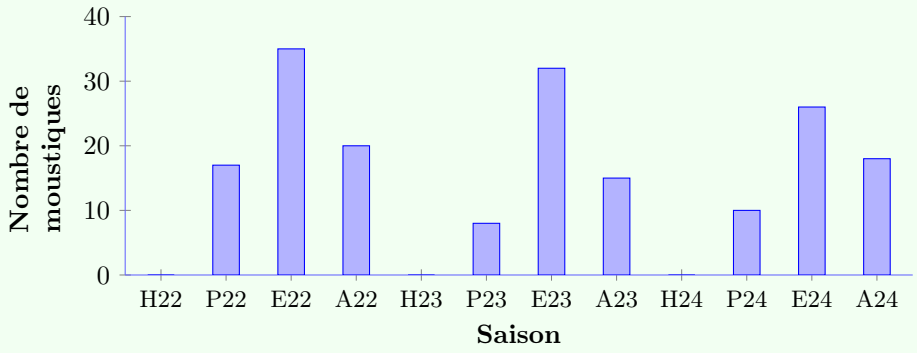
FIGURE IV.2 – Âge des spectateurs dans une salle de cinéma.



Dans cette figure, bien qu'une des valeurs soit clairement plus représentée que le reste, on peut raisonnablement identifier deux modes : la catégorie 0-9 ans (enfants) et la catégorie 30-39 ans (parents) : c'est probablement un film pour enfants qui sont accompagnés de leurs parents. La distribution est donc *bimodale*, reflétant la présence de deux groupes d'âge distincts.

Exemple IV.1.4 : Distribution multimodale

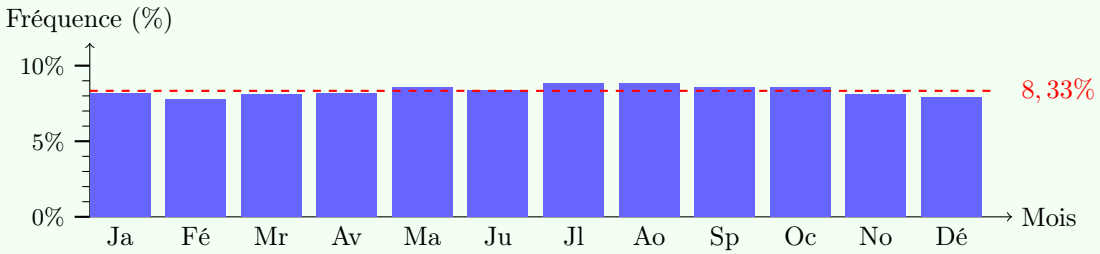
FIGURE IV.3 – Nombre de moustiques écrasés par saison.



On voit dans cette distribution que tous les ans, la saison estivale (E) présente un pic marqué dans le nombre de moustiques écrasés, tandis que les saisons hivernales (H) montrent des creux. La distribution est donc *multimodale*, avec plusieurs pics récurrents chaque année correspondant aux saisons chaudes.

Exemple IV.1.5 : Distribution amodale

FIGURE IV.4 – Proportion des naissances en fonction du mois de l'année.



À titre indicatif, on représente en rouge la fréquence $1/12 = 8{,}33\%$. Dans l'exemple de la figure IV.4, la distribution est techniquement unimodale et le mode est le mois de juillet, qui a la fréquence la plus élevée. Cependant, la différence de fréquence entre juillet et les autres mois est relativement faible, et la distribution est assez plate. On peut donc considérer que cette distribution est *amodale*, car il n'y a pas de pic prononcé indiquant une préférence marquée pour un mois particulier.

Propriété IV.1.6 : Propriétés du mode :

- Applicable à tous les types de variables (nominales, ordinales, quantitatives)
- Peut ne pas exister ou ne pas être unique
- Utile pour identifier les catégories ou valeurs les plus courantes

Interprétation du mode

Le mode est particulièrement utile pour les variables nominales, où les catégories n'ont pas d'ordre intrinsèque. Par exemple, dans une enquête sur la couleur préférée des voitures, le mode indiquerait la couleur la plus populaire parmi les répondants. De plus, le mode peut être utilisé pour identifier des tendances ou des préférences dans des ensembles de données qualitatives, comme les types de produits les plus vendus dans un magasin ou les destinations de voyage les plus populaires.

Dire que le mode d'une distribution de données (x_i) est M_o s'interprète comme :

"La valeur M_o est la plus fréquente dans les données x_i"

Ceci est l'interprétation la plus basique possible. Ensuite, en fonction de valeurs de mode et du contexte des données, on peut en tirer des informations supplémentaires, comme on l'a fait pour l'âge des spectateurs dans la figure IV.2.

IV.1.2 Médiane

Pour les variables utilisant une échelle ordinale ou quantitative, on peut définir une autre manière de mesurer la tendance centrale : la médiane. L'idée est qu'on est au milieu s'il y a autant de valeurs en dessous qu'au-dessus, ce qui explique qu'on ne puisse pas l'appliquer à des variables nominales : on ne peut pas définir ce que veut dire "au-dessus" et "en dessous" pour des catégories sans ordre.

Définition IV.1.7 : Médiane

La médiane est la valeur qui sépare la distribution ordonnée en deux parties égales : 50% des observations sont inférieures et 50% sont supérieures.

Méthode IV.1.8 : Calcul de la médiane

Si on a une série de n données nommées x_i, pour $i = 1, 2, \ldots, n$, la médiane se calcule de la façon suivante :

1. Ordonner les données du plus petit au plus grand : $x_1 \leq x_2 \leq \ldots \leq x_n$
2. Si n est impair : Médiane $= x_{(n+1)/2}$
3. Si n est pair : Médiane $= \frac{x_{n/2}+x_{(n/2)+1}}{2}$.

Attention : On utilise bien $n+1$ et pas n dans la formule pour les données impaires. En effet, si on a 5 données, la médiane est la 3^{e} valeur (et non pas la 2^{e} ou la 4^{e}), ce qui correspond à $n+1$ divisé par 2.

En d'autres termes, si on a un nombre impair de données, la médiane est la valeur centrale une fois les données ordonnées. Si on a un nombre pair de données, la médiane est la moyenne des deux valeurs centrales.

Exemple IV.1.9

Pour les notes d'un examen : 65, 73, 68, 85, 70, 78.

1. On ordonne les notes : 65, 68, 70, 73, 78, 85
2. Il y a 6 notes (pair), donc la médiane est la moyenne des 3^{e} et 4^{e} notes :

$$\text{Médiane} = \frac{70+73}{2} = 71.5$$

Pour les âges d'un groupe d'amis : 25, 27, 35, 30, 25.

1. On ordonne les âges : 25, 25, 27, 30, 35.
2. Il y a 5 âges (impair), donc la médiane est la 3^{e} valeur :

$$\text{Médiane} = 27$$

Propriété IV.1.10 : Propriétés de la médiane :

— Robuste face aux valeurs extrêmes
— Mesure de position centrale appropriée pour les distributions asymétriques
— Applicable aux variables ordinales et quantitatives

Ce que l'on entend par "robuste face aux valeurs extrêmes" est que la médiane n'est pas affectée par des valeurs très élevées ou très basses. Par exemple, si dans un échantillon de 100 personnes, 99 ont un revenu de 50 000 $ et une personne a un revenu de 1 000 000 $, la médiane sera toujours de 50 000 $, car la moitié des personnes gagnent moins que cela et l'autre moitié gagne plus. En revanche, la moyenne serait fortement influencée par le revenu élevé de cette seule personne.

Interprétation de la médiane

La médiane peut être interprétée comme le "point milieu" d'une distribution de données. Elle divise l'ensemble des observations en deux moitiés égales, ce qui en fait une mesure particulièrement utile pour comprendre la répartition des données, surtout lorsqu'il y a des

valeurs extrêmes ou une asymétrie dans la distribution. Par exemple, dans le contexte des revenus, la médiane donne une idée plus précise du revenu "typique" d'une population, car elle n'est pas influencée par les très hauts revenus qui pourraient fausser la moyenne. Par ailleurs, contrairement au mode ou à la moyenne, par définition de la médiane, on sait toujours qu'au moins 50% des données sont en dessous et 50% au-dessus de cette valeur, ce qui est très utile si on veut par exemple mettre en place une politique sociale dont les détails dépendent du nombre de gens dont il faut s'occuper.

Exemple IV.1.11

(Exemple fictif) Supposons qu'une université ait un budget de 1 000 000$ à distribuer pour soutenir les 1000 étudiants d'un certain programme. Si l'université décide de répartir son budget également entre tous les étudiants dont le revenu familial est inférieur à la moyenne des étudiants, il se peut qu'une première année la moyenne soit proche de la médiane et que chaque étudiant éligible reçoive environ 2000$. Cependant, si l'année suivante, un petit nombre d'étudiants très riches s'inscrit dans le programme, la moyenne pourrait augmenter considérablement, au point où 90% des étudiants ont un revenu familial inférieur à la moyenne, ce qui donne une bourse de seulement 1111$ par étudiant éligible. Ainsi, si on se base sur la moyenne, les étudiants moins fortunés recevraient moins d'aide juste parce qu'il y a un petit nombre d'étudiants très riches inscrit dans le programme.
Pour éviter cela, l'université décide de donner des bourses au 50% des étudiants les moins fortunés, ce qui garantit que chaque étudiant éligible recevra une bourse de 2000$ chaque année. Cependant, chaque étudiant individuel ne peut pas savoir s'il est éligible : il ne connait que son propre revenu familial, pas celui des autres étudiants. Ainsi, l'université décide de communiquer que tous les étudiants dont le revenu familial est inférieur à la médiane recevront la bourse [a]. Cela permet à chaque étudiant de savoir s'il est éligible ou non, sans révéler les revenus des autres étudiants. De plus comme la valeur de la médiane ne dépend pas de revenus de quelques étudiants très riches, le nombre de boursiers (et donc, le montant de la bourse) reste stable.

a. Pour information, le revenu familial médian après impôts par personne au Québec en 2022 était de 39 000$.

En pratique, dire que la médiane d'une distribution de données (x_i) est M_d s'interprète comme :

"Au moins la moitié des données x_i sont inférieures ou égales à M_d"

ou, de façon équivalente :

"Au moins la moitié des données x_i sont supérieures ou égales à M_d".

Comme avant, en fonction de la valeur de la médiane et du contexte des données, on peut en tirer des informations supplémentaires. Par exemple, si la médiane est très basse par rapport à la moyenne, cela peut indiquer que la distribution est fortement asymétrique avec une longue queue à droite, ce qui est souvent le cas pour les revenus.

Exemple IV.1.12

La médiane de revenus individuels après impôts au Québec en 2022 est de 39 000$ [a]. Cela signifie qu'en 2022, au moins 50% des individus au Québec avaient un revenu après impôts inférieur ou égal à 39 000$. En d'autres termes, en 2022, 1 personne sur 2 au Québec gagnait moins de 39 000$ par an après impôts. Par contraste, la même source indique que le revenu moyen individuel par personne en 2022 était de 44 500$: la majorité des gens gagnent moins que la moyenne.

a. Source : Statistique Canada, *Enquête canadienne sur le revenu* (2012-2022)

IV.1.3 Moyenne

Si les données sont quantitatives, et donc sont mesurées sur une échelle d'intervalle ou de ratio, on peut définir une autre mesure de tendance centrale : la moyenne.

Définition IV.1.13

La moyenne d'une série de n données $x_1, x_2, \ldots, x_n$ est la somme de toutes les valeurs divisée par le nombre total d'observations.

$$\frac{1}{n}\sum x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Le symbole $\sum$ (sigma) représente l'opération de sommation : $\sum x_i$ se lit "la somme de toutes les valeurs x_i".
On note la moyenne de la population par la lettre grecque μ (mu) et la moyenne de l'échantillon par $\bar{x}$ (x-barre).

Moyenne de la population :

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Moyenne de l'échantillon :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Pour distinguer les deux, on appelle souvent *moyenne empirique* ou *expérimentale* la moyenne de l'échantillon $\bar{x}$, et *moyenne théorique* la moyenne de la population μ.

Il existe d'autres quantités que l'on appelle aussi "moyenne" comme la moyenne *géométrique*. Cependant, par défaut, le terme "moyenne" fait référence à la moyenne *arithmétique*

définie ci-dessus.

Exemple IV.1.14

Pour les notes d'un examen : 65, 72, 68, 85, 70, 78

$$\bar{x} = \frac{65 + 72 + 68 + 85 + 70 + 78}{6} = \frac{438}{6} = 73$$

On peut interpréter la moyenne comme le "centre de gravité" de la distribution des données.

FIGURE IV.5 – La moyenne comme centre de gravité : les données sont en équilibre autour de $\bar{x}$

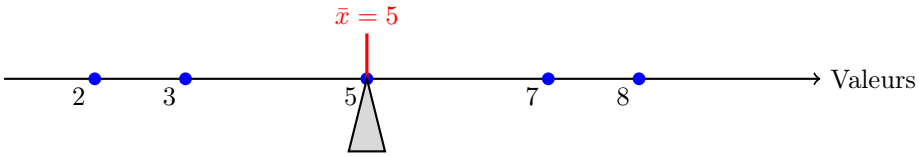
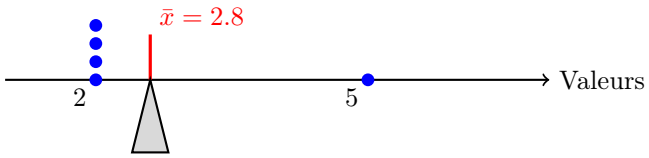


FIGURE IV.6 – La moyenne prend en compte la fréquence des valeurs



Propriété IV.1.15 : (À ne pas retenir, pour votre culture générale.)

La somme des distances signées entre chaque valeur et la moyenne est toujours nulle :

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Démonstration. En effet,

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x}) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} \\ &= \sum_{i=1}^{n} x_i - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0\end{aligned}$$

Ce qu'il fallait démontrer. □

Imaginons que vous posez un bâton en équilibre sur votre doigt. Si vous mettez un poids sur le bâton, le poids fera tourner le bâton avec une intensité proportionnelle à sa masse et à

la distance avec votre doigt (c'est pour cela qu'il est plus facile de porter un sac à dos près du dos qu'à bout de bras : le poids est le même, mais la distance est plus longue). En physique, on appelle cela le *couple*. C'est au sens de la propriété précédente que la moyenne est le point d'équilibre : si on place une masse sur un bâton pour chaque donnée x_i la moyenne est le point où le bâton est en équilibre : le couple total des poids à gauche de la moyenne est égal au couple total des poids à droite de la moyenne.

Il se peut que les données soient regroupées en classes en fonction de la valeur de la variable dont on veut calculer la moyenne, chaque classe ayant une certaine taille. Dans ce cas, la moyenne se calcule en pondérant chaque valeur par sa fréquence.

Exemple IV.1.16

(Données inventées) On a interrogé un millier de personnes sur le nombre de voitures dans leur foyer. Les résultats sont les suivants :

FIGURE IV.7 – Répartition des foyers selon le nombre de voitures

Nombre de voitures	Nombre de foyers	Fréquence	Pourcentage
0	150	0,15	15%
1	400	0,40	40%
2	300	0,30	30%
3	100	0,10	10%
4	40	0,04	4%
5	10	0,01	1%
Total	**1000**	**1,00**	**100**

On peut calculer la moyenne du nombre de voitures par foyer comme suit, à partir des effectifs de chaque classe :

$$\begin{aligned}\bar{x} &= \frac{(0\times 150)+(1\times 400)+(2\times 300)+(3\times 100)+(4\times 40)+(5\times 10)}{1000}\\ &= \frac{0+400+600+300+160+50}{1000} = \frac{1510}{1000} = 1.51\end{aligned}$$

À partir des fréquences relatives, le calcul devient :

$$\begin{aligned}\bar{x} &= (0\times 0,15)+(1\times 0,40)+(2\times 0,30)+(3\times 0,10)+(4\times 0,04)+(5\times 0,01)\\ &= 0+0,40+0,60+0,30+0,16+0,05 = 1,51\end{aligned}$$

Enfin, si on fait le calcul à partir des pourcentages, il ne faut pas oublier de diviser par

100 à la fin :

$$\bar{x} = \frac{(0 \times 15) + (1 \times 40) + (2 \times 30) + (3 \times 10) + (4 \times 4) + (5 \times 1)}{100}$$
$$= \frac{0 + 40 + 60 + 30 + 16 + 5}{100} = \frac{151}{100} = 1,51$$

Il est rassurant de voir qu'on n'a pas cassé les maths et que les trois méthodes donnent le même résultat.

Cet exemple est en fait général. Non seulement on peut calculer la moyenne à partir des distributions de fréquences (que ce soit les fréquences brutes ou relatives), mais c'est en fait en général beaucoup plus pratique de faire ainsi dans les cas (très courants) où on a un grand nombre de données et un petit nombre de valeurs distinctes.

Propriété IV.1.17

La mesure d'une variable dans un groupe de n unités statistiques donne une série de k valeurs $x_1, x_2, \ldots, x_k$ avec des effectifs respectifs $n_1, n_2, \ldots, n_k$ de sorte que $n = \sum_{i=1}^{k} n_i = n_1 + n_2 + \cdots + n_k$. On rappelle que la fréquence relative de la valeur x_i est $f_i = \frac{n_i}{n}$. La moyenne $\bar{x}$ de la série est donnée par les formules équivalentes suivantes :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i \qquad \text{et} \qquad \bar{x} = \sum_{i=1}^{k} f_i x_i$$

Notez qu'on a noté ici $\bar{x}$ la moyenne, donc on a un échantillon, mais la propriété reste valable pour la moyenne de la population μ en remplaçant n par N et $\bar{x}$ par μ.

Propriété IV.1.18 : Propriétés de la moyenne :

- Sensible aux valeurs extrêmes
- Utilise toutes les observations
- Centre de gravité de la distribution
- Applicable uniquement aux variables quantitatives

Interprétation de la moyenne

La moyenne répond à la question : "Si la totalité de la quantité mesurée était répartie également entre toutes les unités statistiques, quelle serait la valeur pour chaque unité ?" Par exemple, si on a un total de 1000 $ réparti entre 10 personnes, la moyenne est de 100 $. Cela signifie que si on redistribuait l'argent de manière égale, chaque personne recevrait 100 $. La moyenne est donc une mesure de tendance centrale qui reflète la répartition globale des

valeurs dans un ensemble de données. À cause de cela, la sensibilité aux valeurs extrêmes est une caractéristique de la moyenne à garder en tête : veut-on une image fidèle de la répartition globale, ou veut-on une image plus "typique" de la majorité des valeurs ?

On a déjà discuté ce que signifie la sensibilité aux valeurs extrêmes : si Jeff Bezos rentre dans une salle de classe, la moyenne des patrimoines des gens dans la classe augmente immédiatement pour atteindre plusieurs milliards de dollars. Pourtant, la richesse de tous les gens qui étaient déjà là n'a pas changé. Inversement, si l'on calcule le nombre moyen de jambes par personne, on va trouver un nombre légèrement inférieur à 2, car la plupart des gens ont 2 jambes, mais certaines personnes en ont moins (amputations, malformations, etc.). Cependant, il est difficile de donner une interprétation claire à cette moyenne, car la notion de fraction de jambe n'existe pas vraiment. Il est encore moins naturel de se poser la question "si on récoltait toutes les jambes pour les répartir de façon égale, combien de jambes aurait chaque personne ?". Pour parler du nombre "typique" de jambes par personne, il vaut mieux utiliser la médiane (2 jambes) que la moyenne.

Cependant, même dans des cas où l'idée de partager une quantité n'a pas de sens évident, la moyenne peut avoir une grande utilité : par exemple, imaginons que l'on veuille créer un service de cardiologie dans une ville de 50 000 habitants qui en était dépourvue jusqu'à présent. On sait qu'en moyenne, chaque habitant a besoin de 0,08 visite cardiologique par an (soit 8 visites pour 100 habitants). En utilisant la moyenne, on peut estimer que la population totale de la ville aura besoin de $50\,000 \times 0{,}08 = 4000$ visites cardiologiques par an et dimensionner le service de façon appropriée. Même si aucun individu ne fait exactement 0,08 visite par an (c'est impossible : certains n'en auront pas du tout, d'autres en auront plusieurs), cette moyenne permet de planifier les ressources nécessaires pour répondre aux besoins de la population. À l'inverse, la médiane et le mode de cette distribution (nombre de visites par habitant par an) sont 0, car la majorité des gens n'ont pas besoin de visite cardiologique chaque année. Si on se basait sur la médiane, on pourrait conclure qu'il n'y a pas besoin de service de cardiologie du tout, ce qui serait manifestement une erreur.

IV.1.4 Comparaison et choix de la mesure appropriée

En fonction de la distribution, la moyenne peut être supérieure ou inférieure à la médiane, ce qui traduit une asymétrie dans la distribution des données. En général, la moyenne est plus affectée par les valeurs extrêmes que la médiane, ce qui peut entraîner une distorsion de l'image de la tendance centrale si la distribution est fortement asymétrique ou contient des outliers. Le mode, quant à lui, peut être très différent de la moyenne et de la médiane, surtout dans les distributions multimodales ou amodales.

Définition IV.1.19

On dit qu'une distribution est :

- **Symétrique** si la moyenne et la médiane sont égaux ou très proches les uns des autres.
- **Asymétrique positive** (ou à droite) si la moyenne est supérieure à la médiane. Cela indique que la distribution a une longue queue à droite.
- **Asymétrique négative** (ou à gauche) si la moyenne est inférieure à la médiane. Cela indique que la distribution a une longue queue à gauche.

FIGURE IV.8 – Dans une distribution symétrique, la moyenne et la médiane coïncident

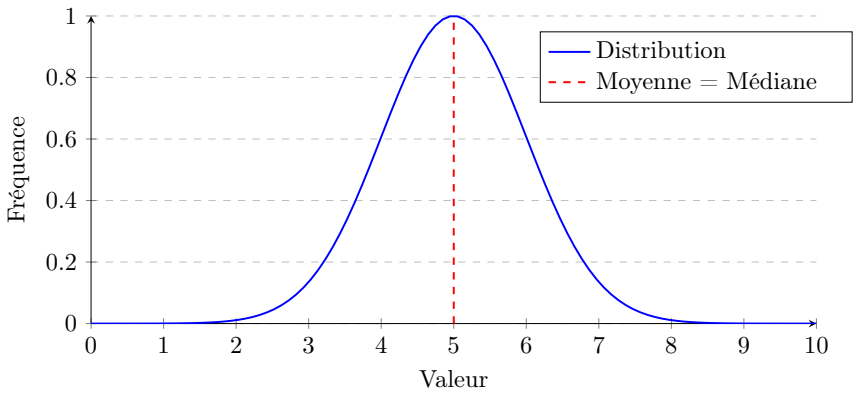


FIGURE IV.9 – Asymétrie positive : Médiane < Moyenne.

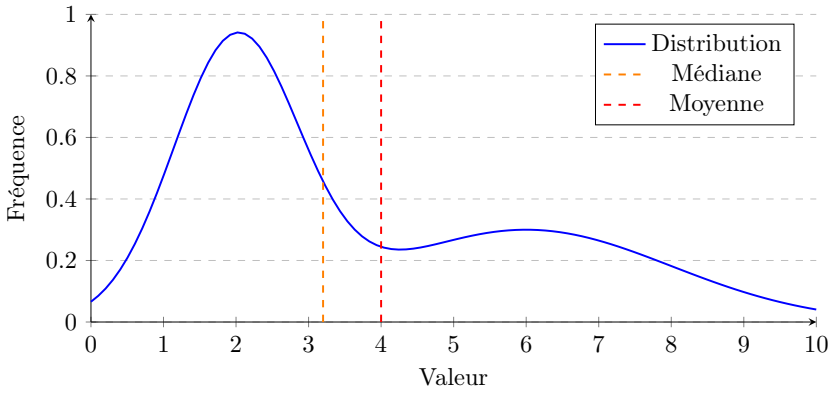
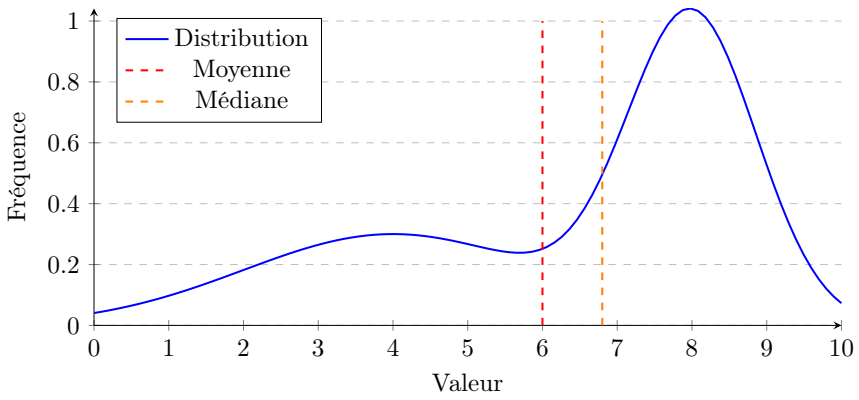


FIGURE IV.10 – Asymétrie négative : Moyenne < Médiane.



Selon la forme de la distribution et la question d'intérêt, l'un ou l'autre des mesures de tendance centrale peut être plus appropriée. Par exemple, pour une distribution symétrique, la moyenne est souvent utilisée car elle est plus facile à manipuler mathématiquement et a des propriétés algébriques utiles. Cependant, pour une distribution asymétrique ou en présence de valeurs extrêmes, la médiane peut être une meilleure mesure de tendance centrale car elle est plus représentative de la "typicalité" des données. Le mode est particulièrement utile pour les variables nominales ou pour identifier les catégories les plus fréquentes dans une distribution.

TABLE IV.1 – Comparaison des mesures de tendance centrale

Situation	**Mesure recommandée**
Distribution symétrique	Moyenne (toutes les mesures sont similaires)
Distribution asymétrique	Médiane (plus représentative)
Présence de valeurs extrêmes	Médiane (robuste)
Variables nominales	Mode (seule mesure applicable)
Variables ordinales	Médiane ou mode
Analyses mathématiques	Moyenne (propriétés algébriques)

IV.2 Mesures de dispersion

Les mesures de tendance centrale décrivent où se situe le "centre" des données, mais elles ne donnent pas d'information sur la façon dont les données sont réparties autour de ce centre. Pour cela, on utilise des mesures de dispersion qui quantifient l'étalement des données. On

ne peut appliquer des mesures de dispersion qu'aux variables quantitatives, car les variables nominales et ordinales n'ont pas de sens en termes de distance entre les valeurs.

IV.2.1 Minimum, maximum, étendue

Définition IV.2.0 : Minimum et maximum

Le minimum d'une série de données est la plus petite valeur de la série, tandis que le maximum est la plus grande valeur. On les note respectivement $\min\{x_i | i = 1, ..., n\}$ et $\max\{x_i | i = 1, ..., n\}$. Bien sûr, si on a affaire à un recensement au lieu d'un échantillon, on peut remplacer n par N dans les notations.

Le maximum et le minimum sont eux-mêmes des mesures intéressantes de la série de données car elles permettent de la situer. Cependant, en termes de dispersion, avec le minimum et le maximum, on peut calculer l'étendue de la série de données :

Définition IV.2.1 : Étendue

L'étendue d'une série de données est la différence entre la valeur maximale et la valeur minimale :

$$\text{Étendue} = \max\{x_i | i = 1, ..., n\} - \min\{x_i | i = 1, ..., n\}.$$

La notation pour l'étendue n'est pas aussi standardisée que pour les mesures suivantes, mais elle est souvent notée E ou R (pour "range" en anglais). Comme avant, si on a affaire à un recensement au lieu d'un échantillon, on peut remplacer n par N dans la formule.

La mesure de l'étendue est très simple à calculer et à comprendre, mais elle transmet peu d'information sur la distribution des données, car elle ne prend en compte que les valeurs extrêmes. Par exemple, si on a les données suivantes : 1, 2, 3, 4, 5, l'étendue est de 4 (5 - 1). Si on ajoute une valeur extrême à ces données, par exemple 100, l'étendue devient 99 (100 - 1), même si la majorité des données sont toujours concentrées entre 1 et 5.

Interprétation de l'étendue

L'étendue peut être interprétée comme la "largeur" de la distribution des données. Elle indique la taille de l'intervalle dans lequel se trouvent toutes les observations, du minimum au maximum. Cependant, elle ne donne pas d'information ni sur la position de l'intervalle, ni sur la répartition des données à l'intérieur de cet intervalle. Si on connait une donnée x et qu'on sait que l'étendue est E, on sait que les données se trouvent dans l'intervalle $[x - E, x + E]$. L'étendue est importante dans les situations où on accorde une importance particulière aux valeurs extrêmes, comme dans les applications où la vie ou la sécurité sont en jeu. Par exemple,

si vous construisez un ascenseur, vous devez vous assurer que la charge maximale supportée par l'ascenseur est suffisante pour supporter le poids de tous les passagers, et pas seulement du passager moyen. Elle est d'autant plus utile que les données sont proches d'avoir une distribution uniforme, c'est-à-dire que les données sont réparties de manière relativement égale entre la valeur minimale et la valeur maximale. Dans ce cas, l'étendue peut donner une bonne indication de la dispersion des données.

IV.2.2 Écart moyen

Définition IV.2.2 : Écart moyen

L'écart moyen d'une série de données est la moyenne des distances absolues entre chaque valeur et la moyenne de la série : $s_1 = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$.

En d'autres termes, c'est la moyenne de l'écart absolu de chaque donnée par rapport à la moyenne. Si on a un recensement, on remplace n par N et $\bar{x}$ par μ dans la formule.

Bien que cette mesure soit une idée "naturelle" de dispersion, elle est rarement utilisée en pratique, car elle ne possède pas de propriétés mathématiques aussi utiles que la variance et l'écart-type. Elle est donc assez peu utilisée. En termes d'interprétation, plus elle est grande, plus on peut considérer que les données sont dispersées autour de la moyenne. Contrairement à la variance et à l'écart-type que nous allons voir ensuite, l'écart moyen accorde la même "importance" à tous les écarts : la série de données

$$10, 0, 0, 0, 0, 0, 0, 0, 0, 0$$

a un écart moyen de 1.8, de même que la série

$$4, 0, 4, 0, 4, 0, 4, 0, 3, 1$$

mais intuitivement la seconde est plus "resserrée" autour de sa moyenne que la première.

IV.2.3 Variance et écart-type

Définition IV.2.3 : Variance

Si on a affaire à une population de taille N et de moyenne μ, la variance de la population est définie par :

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Si on a affaire à un échantillon de taille n et de moyenne $\bar{x}$, la variance de l'échantillon

est définie par :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Notez, et c'est très important, que la formule de la variance de l'échantillon utilise $n-1$ au lieu de n dans le dénominateur. C'est ce qu'on appelle la *correction de Bessel*, et elle est nécessaire pour obtenir une estimation non biaisée de la variance de la population à partir d'un échantillon. En effet, si on utilisait n au lieu de $n-1$, on sous-estimerait systématiquement la variance de la population, surtout pour les petits échantillons. On en fera un exemple en labo.

Définition IV.2.4 : Écart-type

L'écart-type est la racine carrée de la variance :

$$\sigma = \sqrt{\sigma^2} \qquad \text{et} \qquad s = \sqrt{s^2}$$

L'écart-type est plus facile à interpréter que la variance, car il est exprimé dans les mêmes unités que les données d'origine. Par exemple, si les données sont des poids en kilogrammes, l'écart-type sera également en kilogrammes, ce qui facilite la compréhension de la dispersion des données par rapport à la moyenne.

Que représente l'écart-type ? Vous savez que la distance entre deux points (x_1, y_1) et (x_2, y_2) dans un plan est donnée par le théorème de Pythagore :

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Il n'y a pas de raison de se contenter de deux coordonnées et on peut mesurer la distance entre la distribution observée des x_i et la distribution constante égale à la moyenne :

$$d = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{ou, selon le cas, } N, \mu.$$

Cependant, intuitivement, on a envie de dire que l'ensemble de données $(0, 20, 20, 0)$ est plus loin de sa moyenne (10) que l'ensemble de données $(9, 11, 11, 9, 9, ..., 11)$ (200 fois 9 et 200 fois 11). Pourtant, si on fait le calcul pour la première série de données, on trouve :

$$d = \sqrt{(0-10)^2 + (20-10)^2 + (20-10)^2 + (0-10)^2} = \sqrt{400} = 20.$$

Pour la deuxième série de données, on trouve :

$$d = \sqrt{200 \times (9-10)^2 + 200 \times (11-10)^2} = \sqrt{400} = 20.$$

On voit que les deux séries de données sont à la même distance "brute" de leur moyenne. Pourtant, intuitivement, les 9 et 11 sont tous plus proches de la moyenne que les 0 et 20. C'est pour cela qu'on divise par $n-1$ ou N : pour normaliser la distance entre les données et la moyenne et la rendre indépendante du nombre d'observations, plutôt que de calculer la distance totale.

Une autre manière de voir est de dire que le carré dans la formule pénalise beaucoup plus les écarts importants que les écarts faibles. Par exemple, un écart de 10 contribue à la variance pour 100 (10 au carré), tandis qu'un écart de 1 ne contribue que pour 1 (1 au carré) car un grand écart nous éloigne plus de la moyenne que plusieurs petits écarts.

Interprétation de l'écart-type

L'écart-type mesure la concentration des données autour de la moyenne. Plus l'écart-type est faible, plus les données sont proches de la moyenne. Inversement, un écart-type élevé indique que les données sont dispersées autour de la moyenne. Ce qui est toujours vrai, c'est que si les données ont une moyenne μ et un écart-type σ, alors au moins 75% (3/4) des données se trouvent dans l'intervalle $[\mu - 2\sigma, \mu + 2\sigma]$, au moins 89% (8/9) des données se trouvent dans l'intervalle $[\mu - 3\sigma, \mu + 3\sigma]$, et au moins 93,75% (15/16) des données se trouvent dans l'intervalle $[\mu - 4\sigma, \mu + 4\sigma]$, etc. On appelle cela l'*inégalité de Bienaymé-Chebychev*.

Si on sait plus de choses sur la distribution des données, on peut renforcer ces estimations. Par exemple, dans le cas d'une distribution normale, que l'on étudiera plus précisément plus loin, on a la règle suivante : environ 68% des données se trouvent dans l'intervalle $[\mu-\sigma, \mu+\sigma]$, environ 95% des données se trouvent dans l'intervalle $[\mu - 2\sigma, \mu + 2\sigma]$, et environ 99,7% des données se trouvent dans l'intervalle $[\mu - 3\sigma, \mu + 3\sigma]$.

Exemple IV.2.5

On reprend les données de l'exemple IV.1.17 sur le nombre de voitures par foyer. On a calculé que la moyenne du nombre de voitures par foyer est de 1,51. On peut calculer l'écart-type de cette distribution à partir des données du tableau :

FIGURE IV.11 – Répartition des foyers selon le nombre de voitures

Nombre de voitures	**Nombre de foyers**	**Fréquence**	**Pourcentage**
0	150	0,15	15%
1	400	0,40	40%
2	300	0,30	30%
3	100	0,10	10%
4	40	0,04	4%
5	10	0,01	1%
Total	**1000**	**1,00**	**100**

On peut calculer la variance de l'échantillon à partir des données du tableau. On calcule d'abord la somme des carrés des écarts à la moyenne :

$$\begin{aligned}&(0-1,51)^2 \cdot 150 + (1-1,51)^2 \cdot 400 + (2-1,51)^2 \cdot 300 \\ &+(3-1,51)^2 \cdot 100 + (4-1,51)^2 \cdot 40 + (5-1,51)^2 \cdot 10 \\ &= 342,0 + 104,0 + 72,0 + 222,0 + 248,0 + 121,8 = \mathbf{1109,8}\end{aligned}$$

Puis pour calculer la variance, on divise par $n-1 = 999$:

$$s^2 = \frac{1109,8}{999} \approx 1,11$$

Enfin, pour calculer l'écart-type, on prend la racine carrée de la variance :

$$s = \sqrt{1,11} \approx 1,05$$

En regardant la distribution dans le tableau, on voit que la majorité de l'écart-type vient des foyers à 0 et 3 voitures : chaque foyer dans ces classes contribue environ 2,25 à la somme des carrés, contre environ 0,25 pour les foyers à 1 ou deux voitures, soit 5 fois moins. Chaque foyer à 4 ou 5 voitures contribue encore plus individuellement, mais il y en a beaucoup moins, donc leur contribution totale est plus faible que celle des foyers à 0 ou 3 voitures.

Comme pour la moyenne, si on a accès à un tableau de fréquences plutôt qu'aux données brutes, on peut calculer la variance et l'écart-type à partir des données du tableau en pondérant chaque écart au carré par la fréquence de la classe correspondante.

Propriété IV.2.6

La mesure d'une variable dans un échantillon de n unités statistiques donne une série de k valeurs $x_1, x_2, \ldots, x_k$ avec des effectifs respectifs $n_1, n_2, \ldots, n_k$ de sorte que $n = \sum_{i=1}^{k} n_i = n_1 + n_2 + \cdots + n_k$. On rappelle que la fréquence relative de la valeur x_i est $f_i = \frac{n_i}{n}$. On calcule la moyenne $\bar{x}$ de la série à partir des données du tableau, puis on calcule l'écart-type s de la série à partir de la formule suivante :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 \qquad \text{et} \qquad s^2 = \frac{n}{n-1}\sum_{i=1}^{k} f_i(x_i - \bar{x})^2.$$

Si on a affaire à un recensement au lieu d'un échantillon, avec N au lieu de n, μ au lieu de $\bar{x}$ et comme avant des effectifs n_i et fréquence relative n_i/N, on a :

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \mu)^2 \qquad \text{et} \qquad \sigma^2 = \sum_{i=1}^{k} f_i(x_i - \mu)^2.$$

Notez une importante différence entre les formules de la moyenne et de l'écart-type à partir des données du tableau : dans le cas d'un échantillon, on divise par $n-1$ pour obtenir la variance. Or, la fréquence relative f_i est déjà divisée par n, donc pour obtenir la variance, on doit remplacer ce n par un $n-1$: on multiplie donc par $n/(n-1)$. Dans le cas d'un recensement, on divise par N pour obtenir la variance, et la fréquence relative est déjà divisée par N, donc on n'a pas besoin de faire de correction de Bessel : on peut simplement utiliser les fréquences relatives dans la formule de la variance.

IV.2.4 Coefficient de dispersion

Si on ne connait pas la moyenne d'une distribution, il est difficile d'apprécier si un écart-type de 10 est grand ou petit par rapport à ce que l'on mesure. Par exemple, si en pesant 1000 éléphants, je trouve une moyenne de 5000 kg et un écart-type de 10 kg, alors je peux conclure que les éléphants sont très similaires en poids. En revanche, si je pèse des souris et que je trouve une moyenne de 15g et un écart-type de 10 g, même si l'écart-type est plus petit en termes absolus, il est en réalité très grand par rapport à la moyenne, ce qui signifie que les souris sont très différentes en poids. C'est pour cela qu'on utilise le coefficient de dispersion, qui est le rapport entre l'écart-type et la moyenne :

Définition IV.2.7 : Coefficient de dispersion

Le coefficient de dispersion d'une série de données est le rapport entre l'écart-type et la moyenne :

$$CD = \frac{\sigma}{\mu} \quad \text{ou} \quad CD = \frac{s}{\bar{x}}$$

On l'appelle aussi *écart-type* relatif ou *coefficient de variation* et on l'exprime généralement en pourcentages.

Attention : le coefficient de dispersion n'est pas défini si la moyenne est nulle, et il peut être trompeur si la moyenne est très proche de zéro, car dans ce cas, même un petit écart-type peut donner lieu à un coefficient de dispersion très élevé.

Il est possible que le coefficient de dispersion soit supérieur à 100%, ce qui signifie que l'écart-type est plus grand que la moyenne, et que les données sont très dispersées par rapport à la moyenne, voire qu'il soit négatif si la moyenne l'est. Par exemple, si on a une moyenne de 10 et un écart-type de 15, le coefficient de dispersion est de 150%, ce qui indique une grande variabilité des données par rapport à la moyenne.

Exemple IV.2.8

Dans l'exemple précédent, le coefficient de dispersion de cette distribution est de $\frac{1{,}05}{1{,}51} \approx 0,69$, ce qui indique que l'écart-type est environ 69% de la moyenne, suggérant une variabilité modérée du nombre de voitures par foyer.

Interprétation du coefficient de dispersion

On se place dans le cas où la moyenne μ de la population est non nulle. La valeur intéressante à interpréter est en fait la valeur absolue[1] du coefficient de dispersion, car le signe du coefficient de dispersion n'a pas de signification particulière : il peut être négatif si la moyenne est négative, et positif si la moyenne est positive, mais cela ne dit rien sur la dispersion des données. En revanche, plus la valeur absolue du coefficient de dispersion est élevée, plus les données sont dispersées par rapport à la moyenne. Par exemple, un coefficient de dispersion de 50% indique que l'écart-type est égal à la moitié de la moyenne, ce qui suggère une variabilité modérée des données. Un coefficient de dispersion de 200% indique que l'écart-type est deux fois plus grand que la moyenne, ce qui suggère une grande variabilité des données.

Le coefficient de dispersion a une application utile pour mesurer la fidélité d'une mesure. Imaginons que l'on veut mesurer un paramètres ρ d'une population et qu'on a deux méthodes d'estimation A et B que l'on applique plusieurs fois chacune pour obtenir des estimations $r_1^A, r_2^A, ..., r_n^A$ et $r_1^B, r_2^B, ..., r_n^B$. On peut alors calculer le coefficient de dispersion pour chaque

1. La valeur absolue d'un nombre x est x si x est positif et $-x$ s'il est négatif. C'est toujours un nombre positif : c'est la "taille" du nombre, indépendamment de son signe.

méthode et comparer leur fidélité : celle des deux qui a le plus petit coefficient de dispersion est la plus fidèle, car elle donne des estimations plus proches les unes des autres.

IV.2.5 Côte z

Là où le coefficient de dispersion permet de comparer la dispersion de différentes distributions, la côte z permet de comparer la position d'une donnée par rapport à la moyenne dans différentes distributions et de répondre par exemple à la question "si on considère les premiers de classes des classes A et B, lequel est le plus au dessus de son groupe ?". Ce n'est pas aussi évident que de calculer sa note moins la moyenne de sa classe : par exemple, si dans les classes A et B la moyenne est de 75% et que les premier ont 100%, mais que dans la classe A, 80% des notes sont entre 70% et 80%, alors que dans la classe B, 80% des notes sont entre 50% et 100%, alors intuitivement, le premier de la classe A est plus au dessus de sa classe que le premier de la classe B, même si les deux ont la même note.

Définition IV.2.9 : Côte z

La côte z d'une donnée x_i est le nombre d'écarts-types que x_i se trouve au-dessus ou en dessous de la moyenne. Dans le cas d'un échantillon, la côte z de x_i est donnée par :

$$z_i = \frac{x_i - \bar{x}}{s}$$

Dans le cas d'une population, la côte z de x_i est donnée par :

$$z_i = \frac{x_i - \mu}{\sigma}$$

Exemple IV.2.10

Si on reprend encore l'exemple des voitures par foyer, on a calculé que la moyenne est de 1,51 et l'écart-type est de 1,05. La côte z d'un foyer qui a 5 voitures est donnée par :

$$z = \frac{5 - 1,51}{1,05} \approx 3,33$$

Cela signifie que ce foyer se trouve à environ 3,33 écarts-types au-dessus de la moyenne du nombre de voitures par foyer dans la population étudiée. Au contraire, la côte z d'un foyer qui n'a pas de voiture est donnée par :

$$z = \frac{0 - 1,51}{1,05} \approx -1,44$$

Interprétation de la côte z

La côte z permet de parler de la position d'une donnée par rapport à la moyenne en termes de nombre d'écarts-types. Comme vous le voyez dans l'exemple ci-dessus, une côte z négative indique que la donnée se trouve en dessous de la moyenne, tandis qu'une côte z positive indique que la donnée se trouve au-dessus de la moyenne. Plus la valeur absolue de la côte z est grande, plus la donnée est éloignée de la moyenne en termes d'écarts-types. Par exemple, si une donnée a une côte z de 2, et qu'on sait que la moyenne est 8,5 et l'écart-type est 1,3, alors on sait que la donnée se trouve à 2 écarts-types au-dessus de la moyenne, soit à $8,5 + 2 \times 1,3 = 11,1$. De même, si une donnée a une côte z de -1,5, alors on sait que la donnée se trouve à 1,5 écarts-types en dessous de la moyenne, soit à $8,5 - 1,5 \times 1,3 = 6,55$. La côte z est particulièrement utile pour comparer des données provenant de distributions différentes, car elle standardise les données en les exprimant en termes d'écarts-types par rapport à leur propre moyenne.

La côte z n'est en général pas une information suffisante pour connaitre la position d'une donnée dans la distribution, car elle ne dit pas quelle proportion des données se trouve au-dessus ou en dessous de cette côte z. Cependant, on verra dans la suite du cours que quand on a une distribution normale, ce qui est souvent le cas, la côte z d'une donnée permet de déterminer exactement quelle proportion des données se trouve au-dessus ou en dessous de cette donnée.

Exemple IV.2.11 : Une application de la côte z : la côte R

Pour déterminer l'admission à l'université, avant 1995, on utilisait la côte z au Québec. Cela a l'avantage de permettre la comparaison de candidats notés par différents enseignants : on ne considère pas la note absolue, mais seulement à quel point on est au-dessus ou en dessous de la moyenne de la classe ou du groupe.
Cependant, si un groupe est plus fort que les autres, il y aura par nécessité des étudiants avec une faible côte z qui risquent de ne pas être admis alors qu'ils auraient pu être premiers de leur classe dans un autre groupe plus faible. C'est pourquoi, depuis 1995, on utilise la côte R, qui transforme la côte z pour prendre en compte la force du groupe de candidats. La côte R d'un candidat ayant une côte z notée Z est donnée par la formule suivante :

$$R = (Z + IFGZ + 5) \times 5$$

où $IFGZ$ est l'indice de force du groupe, qui est calculé à partir des résultats à l'examen ministériel du secondaire et prend ses valeurs dans $[-2, 2]$. Depuis 2017, on ajoute également un terme $IDGZ$ qui prend en compte l'homogénéité du groupe :

$$R = (Z \cdot IDGZ + IFGZ + 5) \times 5$$

IV.3 Mesures de position

On connait déjà une mesure de position : la médiane. C'est une mesure de position au sens où si on connait la médiane, on peut dire si une unité statistique se situe dans la moitié haute ou basse de la distribution. On en a également évoqué deux autres : le minimum et le maximum, qui encadrent la distribution. Cependant, il existe d'autres mesures de position qui permettent de diviser la distribution en plus de deux parties égales, ou qui permettent de classer les données selon leur rang.

IV.3.1 Quantiles

Définition IV.3.0 : Quantiles

Les *quantiles* sont des valeurs qui divisent une distribution ordonnée en parties égales. Le p-ième quantile est la valeur en dessous de laquelle se trouve une proportion p des données. Les quantiles les plus couramment utilisés sont :

- Les **quartiles** (pour *quart*) (Q_1, Q_2, Q_3) : divisent la distribution en quatre parties égales. Q_1 est la plus petite valeur en dessous de laquelle se trouve 25% (1/4) des données, Q_2 est la médiane (50%), et Q_3 est la plus petite valeur en dessous de laquelle se trouve 75% (3/4) des données.
- Les **déciles** $(D_1, D_2, ..., D_9)$: divisent la distribution en dix parties égales. Le i-ième décile est la plus petite valeur en dessous de laquelle se trouve $10i$% des données.
- Les **centiles** $(C_1, C_2, ..., C_{99})$: divisent la distribution en cent parties égales. Le i-ième centile est la plus petite valeur en dessous de laquelle se trouve i% des données.

Légèrement moins fréquemment, on peut aussi trouver des quantiles qui divisent la distribution en 5 parties égales (quintiles, notés V_1, V_2, V_3, V_4, V pour 5 en chiffres romains) ou en 20 parties égales (vingtiles), mais les quartiles, déciles et centiles sont de loin les plus couramment utilisés.

Méthode IV.3.1 : Détermination des quantiles

La méthode générale pour déterminer les quantiles d'une série de données est la suivante (expliquée ici pour un échantillon). Supposons que nous voulons trouver le p-ième quantile (par exemple, Q_3) d'une série de n (par exemple, $n = 10$) données x_i :

1. Ordonner les données du plus petit au plus grand : $x_1 \leq x_2 \leq \ldots \leq x_n$
2. Calculer la position du p-ième quantile par la formule : $k = p \times n$ (par exemple, pour Q_3, $p = 0.75$, donc $k = 0.75 \times 10 = 7.5$).

3. Si k est un entier, alors le p-ième quantile est la moyenne des valeurs à la position k et à la position $k+1$ dans la série ordonnée : $Q_p = \frac{x_k + x_{k+1}}{2}$.
4. Si k n'est pas un entier, alors le p-ième quantile la $\lceil k \rceil$-ième [a] valeur. Dans notre exemple, $k = 7.5$, donc $Q_3 = x_8$.

Attention : Contrairement à la médiane, on utilise bien n dans la formule de la position du quantile, et non pas $n+1$.

a. On rappelle que $\lceil x \rceil$ est la partie entière supérieure de x, autrement dit, son arrondi par excès à un entier.

Exemple IV.3.2

Supposons que nous avons les notes suivantes de 15 étudiants à un examen :

$$45,\ 52,\ 58,\ 61,\ 65,\ 68,\ 72,\ 75,\ 78,\ 82,\ 85,\ 88,\ 91,\ 94,\ 97$$

Calculons le 3$^{\text{e}}$ décile D_3 :

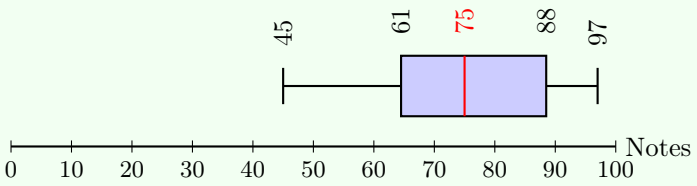
1. Les données sont déjà ordonnées : $x_1 = 45, x_2 = 52, ..., x_{15} = 97$
2. Position du 3$^{\text{e}}$ décile : $k = 0.3 \times 15 = 4.5$
3. Comme $k = 4.5$ n'est pas un entier, on prend la 5$^{\text{ème}}$ valeur : $D_3 = x_5 = 65$

Ainsi, $D_3 = 65$ signifie qu'au moins 30% des étudiants ont obtenu une note inférieure ou égale à 65.

Une représentation des quartiles : la boite à moustache. Une boîte à moustaches (ou diagramme en boîte) représente visuellement les quartiles d'une distribution. La boîte centrale s'étend de Q_1 à Q_3 et contient donc 50% des données. La ligne rouge à l'intérieur représente la médiane Q_2. Les "moustaches" s'étendent jusqu'aux valeurs minimale et maximale, montrant l'étendue complète des données.

Exemple IV.3.3

FIGURE IV.12 – Diagramme en boîte des notes de l'exemple précédent



Typiquement, on n'écrit pas les valeurs des quartiles, minimum, maximum et de la médiane sur le diagramme en boîte, mais on les ajoute ici pour faciliter l'interprétation. Dans ce diagramme, on voit que la majorité des notes se trouvent entre 61 et 88, avec une médiane à 75, ce qui indique que la moitié des étudiants ont obtenu une note inférieure ou égale à 75. Les moustaches montrent que les notes s'étendent de 45 à 97, indiquant une certaine dispersion des résultats.

Une mesure de dispersion autour de la médiane : l'écart interquartile. L'écart interquartile (qu'on note ici EIQ, mais attention, ce n'est pas une notation universelle) est une mesure de dispersion qui indique l'étendue de la partie centrale d'une distribution. Il est calculé en soustrayant le premier quartile Q_1 du troisième quartile Q_3 :

Définition IV.3.4 : Écart interquartile

L'écart interquartile (EIQ) d'une série de données est la différence entre le troisième quartile Q_3 et le premier quartile Q_1 :

$$EIQ = Q_3 - Q_1$$

Contrairement à l'écart-type, qui mesure la dispersion autour de la moyenne, l'écart interquartile mesure la dispersion autour de la médiane. Il est particulièrement utile pour les distributions asymétriques ou contenant des valeurs extrêmes (outliers), car il n'est pas influencé par ces valeurs extrêmes contrairement à l'écart-type.

Il peut être intéressant de comparer l'écart interquartile à l'étendue. L'EIQ est nécessairement plus petit, mais le rapport entre les deux (EIQ/E) est proche de 1, cela veut dire que la distribution est concentrée près de ses valeurs extrêmes, s'il est proche de 1/2 (et que la médiane est au centre), cela veut dire que la distribution est à peu près homogène et s'il est proche de 0, cela veut dire que la distribution est très concentrée autour de sa médiane.

IV.3.2 Rang quantile

Le rang quantile est, d'une certaine façon, l'opposé du quantile : au lieu de connaitre la proportion de données et de chercher la valeur qui les majore, on connait la valeur et on cherche la proportion de données qu'elle majore. Par exemple, si on sait que $Q_3 = 20$, alors le rang quantile de 20 est 0,75, car 75% des données sont inférieures ou égales à 20.

Définition IV.3.5 : Rang quantile

Le rang quantile d'une valeur x dans une série de données est la proportion de données qui sont inférieures ou égales à x. Il est calculé en ordonnant les données et en déterminant la position de x dans cette série ordonnée. Si x se trouve à la position k dans la série ordonnée, alors le rang quantile de x est donné par la formule : Rang quantile $= \frac{k}{n}$, où n est le nombre total de données.

Méthode IV.3.6

La définition elle-même donne la méthode :

1. On ordonne les données de la plus petite à la plus grande : $x_1 \leq x_2 \leq \ldots \leq x_n$
2. On cherche la donnée x : si elle est dans les données, on note k sa position (si elle y est plusieurs fois, on prend la plus grande).
3. Si x n'est pas dans les données, on trouve les deux données x_i et x_{i+1} telles que $x_i < x < x_{i+1}$, et on note $k = i$. (Si x est plus petit que toutes les données, on prend $k = 0$, et si x est plus grand que toutes les données, on prend $k = n$.)
4. Le rang quantile de x est alors donné par la formule : Rang quantile $= \frac{k}{n}$.

IV.3.3 Lire des quantiles

Dans la majorité des cas, on a besoin de calculer les quantiles (y compris la médiane) ou des rangs quantiles à partir de données déjà traitées, dans des tableaux de fréquences ou des graphes.

Lire des quantiles et des rangs dans un tableau de fréquence On rappelle que les quantiles ne peuvent être définis que pour une variable au moins ordinale. Ceci étant dit, si on a une modalité, valeur ou classe v d'une variable ordinale, on peut calculer la fréquence cumulée relative[2] des données inférieures ou égales à v :

$$F_{\leq v} = \sum_{x_i \leq v} f_i$$

où f_i est la fréquence relative de la valeur x_i. Pour trouver le p-ième quantile, il suffit de trouver la première valeur (ou modalité, ou classe) v pour laquelle $F_{\leq v} \geq p$. Par exemple, pour trouver la médiane, il suffit de trouver la première valeur v pour laquelle $F_{\leq v} \geq 0,5 = 50\%$.

Dans le cas d'une variable continue, ou même d'une variable discrète avec un grand nombre de valeurs, il est rare que les tableaux représentent chaque valeur individuellement : dans ce

2. La formule se lit "la fréquence cumulée des modalités/valeurs/classes inférieures ou égales à v, notée $F_{\leq v}$ ici, est la somme des fréquences relatives individuelles des modalités/valeurs/classes inférieures ou égales à v."

cas, on parle de *classe médiane* (ou classe quantile).

Exemple IV.3.7

Reprenons l'exemple du nombre de t-shirts vendus par taille :

Taille	S	M	L	XL
Nombre vendu	4	9	5	2

Il faut d'abord calculer les fréquences relatives et les fréquences relatives cumulées :

Taille	**Effectif**	**Fréquence relative**	**Fréquence relative cumulée (%)**
S	4	0,20	20%
M	9	0,45	65%
L	5	0,25	90%
XL	2	0,10	100%

Pour trouver la médiane, il faut trouver la première classe pour laquelle la fréquence relative cumulée est supérieure ou égale à 50%. Ici, c'est la classe M, donc la médiane est M. Pour trouver le premier quartile (ce qui revient au 25$^{\text{e}}$ centile), il faut trouver la première classe pour laquelle la fréquence relative cumulée est supérieure ou égale à 25%. Ici, c'est aussi la classe M, donc $Q_1 = \text{M}$. De même, pour trouver Q_3, il faut trouver la première classe pour laquelle la fréquence relative cumulée est supérieure ou égale à 75%. Ici, c'est la classe L, donc $Q_3 = \text{L}$.

Inversement, si on a une valeur v et que l'on veut trouver son rang quantile à partir d'un tableau de fréquences cumulées, on a deux cas de figure :

— soit la valeur v est présente dans le tableau, et alors son rang quantile est simplement la fréquence cumulée relative de cette valeur,

— soit la valeur v n'est pas une valeur du tableau, mais on peut trouver v_i et v_{i+1} telles que $v_i < v < v_{i+1}$, et on prend comme rang quantile de v la fréquence cumulée relative de v_i, c'est-à-dire la proportion de données inférieures ou égales à v_i.

— soit on a une valeur v mais il y a des classes dans les lignes. Dans ce cas, on trouve la classe C telle que $v \in C$, et on prend comme rang quantile de v la fréquence cumulée relative de la classe C précédente à C, c'est-à-dire la proportion de données inférieures ou égales à la classe précédente à C.

Exemple IV.3.8

Reprenons l'exemple des véhicules du jeu de données `mtcars` et de leur consommation en miles par gallon (mpg). Supposons que nous avons le tableau de fréquences suivant pour la variable "consommation en mpg" :

TABLE IV.2 – Répartition des 32 véhicules du jeu `mtcars` par consommation en mpg

Consommation (mpg)	**Nombre de véhicules**	**Fréquence cumulée des véhicules (%)**
10-14.99	5	16%
15-19.99	13	56%
20-24.99	8	81%
25-29.99	2	88%
30+	4	100%
Total	32	100%

Si je veux trouver le rang quantile de 23 mpg, je vois que 23 est dans la classe 20-24.99, donc je prends la fréquence cumulée relative de la classe précédente, qui est 56%. Donc le rang quantile de 23 mpg est de 56%, ce qui signifie que 56% des véhicules ont une consommation inférieure ou égale à 23 mpg. En effet, vu que les données sont regroupées en classes, on ne sait pas comment se situent les 8 véhicules de la classe 20-24.99, par rapport à 23 mpg : certains peuvent être en dessous de 23, d'autres au dessus. En prenant la fréquence cumulée relative de la classe précédente, on s'assure de ne pas surestimer le rang quantile de 23 mpg.

Il existe des méthodes qui permettent d'estimer les quantiles et les rangs quantiles à partir de données regroupées en classes, mais elles sont plus complexes et nécessitent des hypothèses supplémentaires sur la distribution des données à l'intérieur de chaque classe. Nous n'aborderons pas ces méthodes dans ce cours, mais il est important de savoir qu'elles existent.

Lire des quantiles et des rangs sur une ogive Puisque la lecture des quantiles et des rangs se fait sur la proportion cumulée, il est naturel que le type de diagramme qui permet de les lire facilement soit l'ogive, qui représente la fréquence cumulée relative en fonction des valeurs de la variable.

Méthode IV.3.9

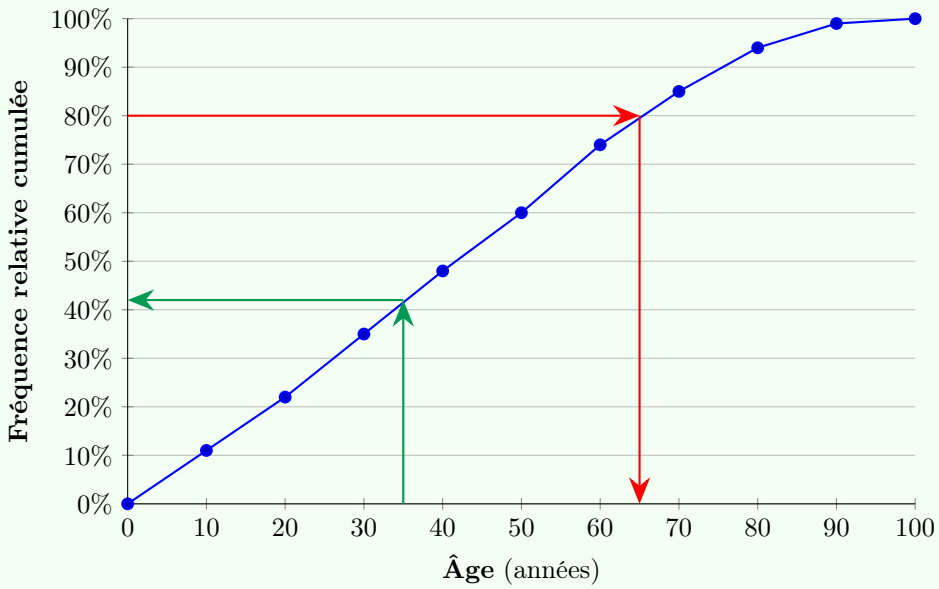
- **Lire un quantile**. Pour lire le p-ième quantile, on trace un trait horizontal à la proportion cumulée relative p sur l'axe des ordonnées. Ce trait croise l'ogive en un

point : la valeur du quantile est l'abscisse de ce point.

• **Lire un rang quantile**. Pour lire le rang quantile d'une valeur v, on trace un trait vertical à la valeur v sur l'axe des abscisses. Ce trait croise l'ogive en un point : le rang quantile de v est l'ordonnée de ce point.

Exemple IV.3.10

FIGURE IV.13 – Distribution cumulative de la population du Québec par groupe d'âge (2021, fréquences relatives cumulées)



On voit (en rouge), que le 0,8$^{\text{ème}}$ quantile (c'est-à-dire, de façon équivalente : le 4$^{\text{ème}}$ quintile, le 8$^{\text{ème}}$ décile ou le 80$^{\text{ème}}$ centile) de la distribution correspond à une valeur d'environ 65 ans, ce qui signifie que 80% de la population du Québec a 65 ans ou moins. De même, on voit (en vert), que le rang quantile de 35 ans est d'environ 42 %, ce qui signifie que 42% de la population du Québec a moins de 35 ans.

Résumé du chapitre

Trois mesures de tendance centrale

Mode	Médiane	Moyenne
Valeur la plus fréquente	Valeur centrale	Somme ÷ nombre
Applicable aux nominales	50% au-dessous, 50% au-dessus	Données quantitatives seulement
Peut être non unique	Robuste aux extrêmes	Sensible aux valeurs extrêmes
Exemple : couleur préférée	Revenu médian	Moyenne salariale

Asymétrie de la distribution

Asymétrie négative	Symétrique	Asymétrie positive
Moy $<$ Méd	Moy $\simeq$ Méd	Moy $>$ Méd

Mesures de tendance centrale : quand utiliser chacune ?

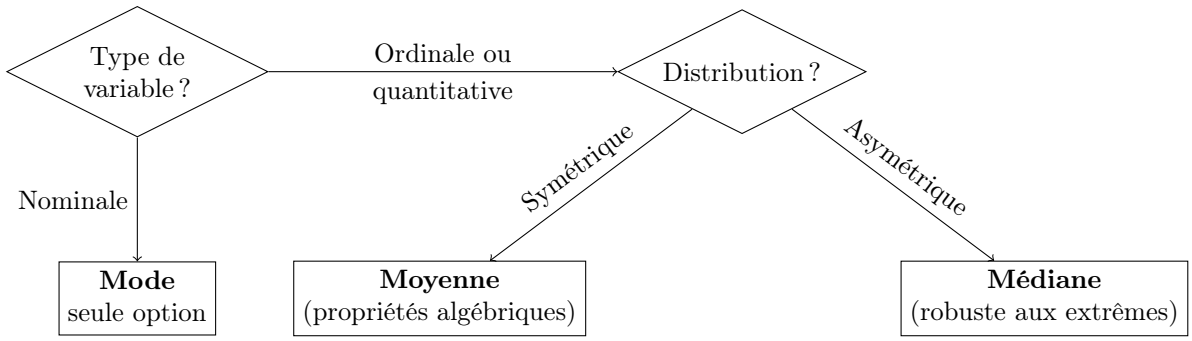


FIGURE IV.14 – Arbre de décision : tendance centrale

Mesures de dispersion : Variabilité des données

Mesure	Interprétation
Étendue	Différence entre max et min (sensible aux extrêmes)
Écart-type	Distance moyenne des données à la moyenne (unités originales)
Variance	Carré de l'écart-type (unités au carré)
Écart interquartile (IQR)	Intervalle contenant les 50% du centre (robuste)

Chapitre V

Statistiques inférentielles

Les statistiques **descriptives**, vues dans le chapitre précédent, permettent de résumer et de visualiser des données. Elles répondent à des questions du type : *quelle est la moyenne des observations de mon échantillon ?* ou *quelle est la dispersion de mes données ?*

Les statistiques **inférentielles** vont plus loin : elles permettent de tirer des conclusions sur une *population entière* à partir d'un *échantillon*. En d'autres termes, on cherche à inférer des propriétés d'un grand ensemble (la population) à partir d'un petit nombre d'observations (l'échantillon).

Exemple V.0.0

Un sondage interroge 1 000 personnes sur leurs intentions de vote. On ne connait pas les intentions de l'ensemble de la population, mais on souhaite les *estimer* à partir de cet échantillon. Les statistiques inférentielles nous donnent les outils pour quantifier la précision de cette estimation.

Ce passage du particulier au général est rendu possible grâce à des outils mathématiques, notamment la **loi normale** et le **théorème central limite**, que nous allons étudier dans ce chapitre.

V.1 Loi normale

V.1.1 Courbe de Gauss

Définition V.1.0 : Courbe de Gauss

Une courbe de Gauss, ou courbe gaussienne, courbe normale, courbe en cloche, normale, est une fonction définie par la formule :

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

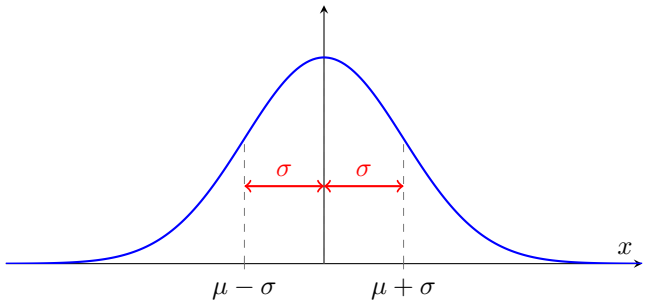
où μ et σ sont des paramètres de la distribution : μ est la moyenne et σ est l'écart-type.

Si la moyenne μ est égale à 0 on dit que la courbe est *centrée*, et si l'écart-type σ est égal à 1 on dit que la courbe est *réduite*. Si les deux conditions sont vérifiées, c'est-à-dire si la courbe est centrée et réduite, on dit que la courbe est *standardisée*. La courbe normale standardisée a donc pour formule :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Notez que vous n'avez pas besoin d'apprendre par cœur la formule de la courbe de Gauss, mais il est important de comprendre les rôles des paramètres μ et σ dans la formule.

FIGURE V.1 – Courbe de Gauss de moyenne μ et d'écart-type σ.



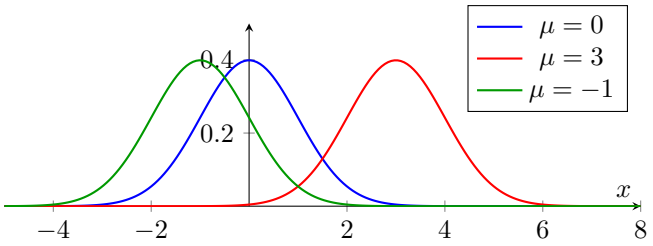
Effets des paramètres μ et σ sur la courbe de Gauss

Les deux paramètres μ et σ contrôlent respectivement la **position** et la **forme** (ou, si vous préférez, la concentration) de la courbe normale.

Effet de la moyenne μ Le paramètre μ détermine le centre de la courbe. Modifier μ translate la courbe horizontalement sans changer sa forme.

On observe que la courbe conserve exactement la même forme, mais se déplace vers la droite lorsque μ augmente. La moyenne μ correspond toujours au sommet de la courbe.

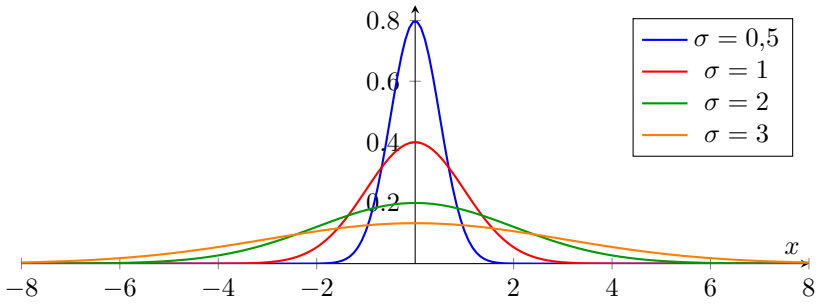
FIGURE V.2 – Décalage de la courbe sous l'effet de μ.



On voit également que la courbe est symétrique par rapport à la verticale passant par μ, où la courbe atteint son maximum : on peut donc lire graphiquement la valeur de μ en lisant la position du sommet de la courbe.

Effet de l'écart-type σ Le paramètre σ contrôle l'étalement de la courbe. Un σ petit produit une courbe haute et étroite, tandis qu'un σ grand produit une courbe basse et étalée. L'aire sous la courbe reste toujours égale à 1.

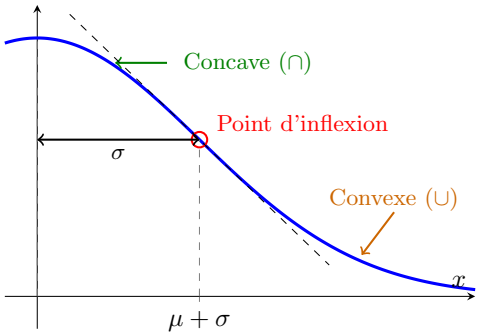
FIGURE V.3 – Étalement de la courbe sous l'effet de σ.



On voit que la courbe devient de plus en plus étalée et de moins en moins haute à mesure que σ augmente.

On peut estimer graphiquement l'écart-type σ en observant la distance entre la moyenne μ et les points d'inflexion de la courbe, c'est-à-dire les points où la courbe change de "courbé vers le bas" près de son maximum à "courbé vers le haut" dans les queues. Ces points d'inflexion se trouvent à une distance de σ de la moyenne μ.

FIGURE V.4 – Point d'inflexion et estimation graphique de σ.



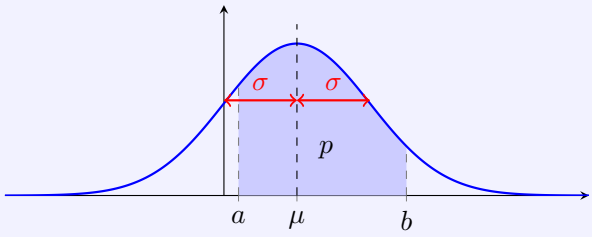
V.1.2 Loi normale

Définition V.1.1 : Loi normale

On dit qu'une variable quantitative x suit une loi normale de paramètres μ et σ si la proportion des observations de x entre deux valeurs a et b (quand le nombre d'observations tend vers l'infini) est égale à l'aire sous la courbe de Gauss de paramètres μ et σ entre les points a et b.

En d'autres termes, si p est la proportion d'observations de x dans l'intervalle $[a, b]$, alors :

$$p = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx =$$



La notation mathématique $\int_a^b$ se lit « intégrale de a à b » et représente l'aire sous une courbe entre les points a et b.

Si x suit une loi normale de paramètres μ et σ, on note $x \sim \mathcal{N}(\mu, \sigma^2)$.

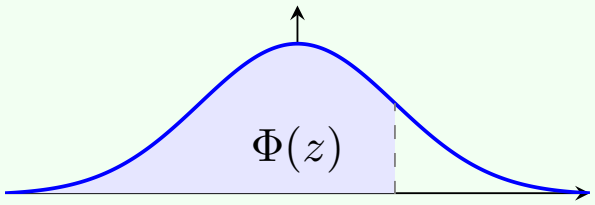
Pour des a et b quelconques, il n'existe malheureusement pas de formule simple pour calculer l'aire sous la courbe. Pour calculer cette aire, on utilise soit des tables (comme vous le ferez à l'examen), soit des ordinateurs (comme vous le ferez en Excel).

Exemple V.1.2

Si on considère une variable z qui suit une loi normale standard, on peut considérer la fréquence cumulée de z à une valeur a donnée, c'est-à-dire la proportion d'observations

de z inférieures ou égales à a. Cette fréquence cumulée correspond à l'aire sous la courbe de Gauss standardisée entre $-\infty$ et a. Le $-\infty$ vient du fait qu'on ne limite pas z vers le bas, on peut avoir des valeurs aussi petites que l'on veut et donc $-\infty$ est la seule chose qui est inférieure à toutes les valeurs possibles de z.
Elle s'appelle la *fonction de répartition* de la loi normale standard et on la note $\Phi(z)$:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx =$$

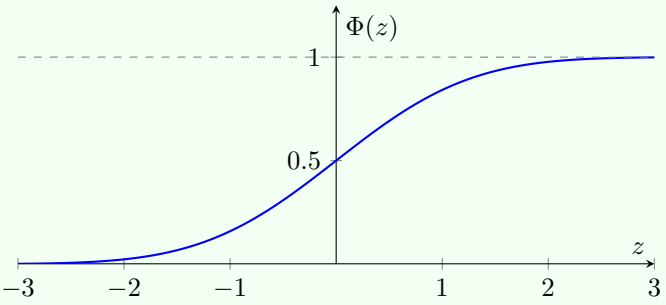


Les valeurs de la fonction $\Phi(z)$ sont comme suit :

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
⋮										
0.9	0.8159	0.8185	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621

Attention, la notation Φ n'est pas standardisée : dans certains livres, on peut trouver la notation $F(z)$ ou $P(z)$ pour la même fonction.
En pratique, même si on ne peut pas la calculer directement sans l'usage d'une table, la fonction Φ a l'allure suivante :

FIGURE V.5 – Graphique de la fonction $\Phi(z)$.



Dans ces conditions, comment calculer à la main la proportion d'observations de x supérieures à 14,86 si on sait que x suit une loi normale de moyenne $\mu = 10$ et d'écart-type

$\sigma = 5$ (par exemple) ? En théorie, il faudrait avoir accès à une table de la courbe de Gauss de moyenne 10 et d'écart-type 5. Il faudrait en fait avoir une table pour chaque valeur de μ et de σ que l'on pourrait rencontrer, ce qui est évidemment impossible. Heureusement, il existe une astuce.

Propriété V.1.3

Si x suit une loi normale de moyenne μ et d'écart-type σ, alors la cote z de x,

$$z = \frac{x - \mu}{\sigma},$$

suit une loi normale standard, c'est-à-dire de moyenne 0 et d'écart-type 1.

Cela nous donne la méthode suivante :

Méthode V.1.4 : Utiliser la table de la loi normale standard

Supposons qu'on a une variable x qui suit une loi normale de moyenne μ (par exemple 10) et d'écart-type σ (par exemple 5), et $a < b$ des nombres (par exemple -5 et 14,86).

Calculer la proportion en dessous de b.

1. On calcule la cote z de b : $z = \dfrac{b - \mu}{\sigma}$. Dans notre exemple, $z = \frac{14{,}86-10}{5} = 0,97$.
2. Dans la table de $\Phi(z)$, on regarde la valeur à l'intersection de la colonne correspondant aux centièmes de z (ici, $0,07$) et de la ligne correspondant aux dixièmes de z (ici, $0,9$). On trouve $\Phi(0,97) = 0,8340$.

Calculer la proportion au-dessus de a.

1. On calcule la cote z comme avant et on lit $\Phi(z)$: la proportion d'observations de x inférieures ou égales à a, comme dans le point précédent.
2. On calcule la proportion d'observations de x supérieures à a :

$$1 - \Phi(z).$$

En effet, une unité statistique est soit au-dessus de a, soit en dessous de a. Donc, pour calculer la proportion d'observations de x au-dessus de a, on peut calculer :

$$\underbrace{\text{Tout le monde}}_{1} \underbrace{-}_{\text{sauf}} \underbrace{\text{ceux qui sont en dessous de } a}_{\Phi(z)} = 1 - \Phi(z).$$

Dans notre exemple, la proportion d'observations de x supérieures à $a = -5$ est égale à $1 - \Phi(-3) = 1 - 0,0013 = 0,9987$.

Calculer la proportion entre a et b. On est entre a et b si on est en dessous de b

mais pas en dessous de a.

1. On calcule z_a et z_b les cotes z de a et de b comme avant et on lit $\Phi(z_a)$ et $\Phi(z_b)$ dans la table de $\Phi(z)$.
2. On calcule la proportion d'observations de x entre a et b :

$$\Phi(z_b) - \Phi(z_a).$$

Dans notre exemple, la proportion d'observations de x entre $a = -5$ et $b = 14,86$ est égale à $\Phi(0,97) - \Phi(-3) = 0,8340 - 0,0013 = 0,8327$.

Comme la fonction de Gauss décroit très rapidement à mesure que l'on s'éloigne de la moyenne μ, les observations de la variable x ayant une cote z très négative ou très positive sont très rares. Dans l'exemple précédent, la proportion d'observations de x inférieures à -5 (c'est-à-dire avec une cote z inférieure ou égale à -3) est de seulement $0,13\,\%$.

Il ne vous faut évidemment pas connaitre par cœur les valeurs de la table de $\Phi(z)$, mais il est important d'en connaitre certaines valeurs spéciales données ci-après, et de savoir que si z est inférieur à -3 ou supérieur à 3, alors $\Phi(z)$ est très proche de 0 ou de 1, respectivement.

FIGURE V.6 – Valeurs remarquables de la fonction cumulée.

Intervalle	**Proportion**
$[\mu - \sigma,\ \mu + \sigma]$	68,27 %
$[\mu - 1,96\sigma,\ \mu + 1,96\sigma]$	95,00 %
$[\mu - 2\sigma,\ \mu + 2\sigma]$	95,45 %
$[\mu - 2,58\sigma,\ \mu + 2,58\sigma]$	99,00 %
$[\mu - 3\sigma,\ \mu + 3\sigma]$	99,73 %
$[\mu - 3,29\sigma,\ \mu + 3,29\sigma]$	99,90 %

V.1.3 Des phénomènes "normaux"

Il est très fréquent que des grandeurs de la vraie vie suivent d'elles-mêmes une loi normale : c'est le cas dès que de nombreux facteurs indépendants contribuent à la grandeur en question. Par exemple, la taille d'une personne est influencée par de nombreux facteurs génétiques et environnementaux, qui sont tous indépendants les uns des autres. C'est pourquoi la taille suit souvent une loi normale. C'est le cas d'autres grandeurs : le poids, le QI, le temps de trajet au travail.

V.2 Estimation d'un paramètre

Pourquoi prendre le temps de discuter d'une répartition en particulier alors que nous avons vu que les distributions de fréquences peuvent prendre des formes très variées ? D'abord – on l'a déjà évoqué – parce que la loi normale est une distribution très fréquente dans les sciences humaines, et que de nombreux phénomènes suivent une loi normale. Ensuite, parce que la loi normale est très importante pour l'estimation d'un paramètre à partir d'un échantillon, ce qui, on le rappelle, est tout l'objectif des statistiques en tant que science. Ces deux raisons sont en fait liées : leur cause commune est un phénomène mathématique appelé le **théorème central limite**, qui explique pourquoi la loi normale est si fréquente et pourquoi elle est si importante pour l'estimation d'un paramètre à partir d'un échantillon.

V.2.1 Théorème central limite

Définition V.2.0 : Observations indépendantes

On dit que des observations sont indépendantes si la réalisation de l'une d'entre elles n'influence pas la réalisation des autres.

Exemple V.2.1

Un lancer de pièce de monnaie n'influence pas les suivants : chaque observation de pile ou de face au cours d'une série de lancers est indépendante de toutes les autres.
Inversement, si on interroge un professeur dans sa classe sur ce qu'il pense des diagrammes en secteurs et qu'il dit qu'ils sont la pire sorte de diagramme, puis qu'on interroge un élève de cette classe sur son opinion sur les diagrammes en secteurs, il est probable que l'élève soit influencé par l'opinion du professeur, et donc que les observations du professeur et de l'élève ne soient pas indépendantes.

Propriété V.2.2 : Théorème central limite

On considère x, une variable quantitative quelconque, suivant une distribution quelconque (et en particulier pas nécessairement une loi normale) de moyenne μ et d'écart-type σ. Soit n un entier positif et $x_1, \ldots, x_n$ des observations indépendantes de x. Alors, quand n devient grand, le résultat du calcul

$$\sqrt{n} \cdot \frac{\bar{x} - \mu}{\sigma}$$

suit une loi normale standard.

C'est à cause de ce théorème que de nombreux phénomènes suivent une loi normale : en effet, si une variable x est influencée par de nombreux facteurs indépendants, alors x peut être

considérée comme la somme de nombreuses variables indépendantes, chacune correspondant à l'influence d'un facteur. Par conséquent, x suit une loi normale.

Exemple V.2.3

Si x a une distribution de moyenne $\mu = 10$ et d'écart-type $\sigma = 5$, (mais ne suit pas nécessairement une loi normale), et $x_1, \ldots, x_{100}$ sont 100 observations indépendantes de x, alors la probabilité que $\bar{x}$ soit entre 9 et 11 est :

$$\begin{array}{rccccc} & 9 & \leq & \bar{x} & \leq & 11 \\ \iff & \frac{9-10}{5} & \leq & \frac{\bar{x}-10}{5} & \leq & \frac{11-10}{5} \\ \iff & -0,2 & \leq & \frac{\bar{x}-10}{5} & \leq & 0,2 \\ \iff & -2 & \leq & \sqrt{100} \cdot \frac{\bar{x}-10}{5} & \leq & 2, \end{array}$$

On voit qu'il y a une environ 95,45% de chances que la moyenne $\bar{x}$ de 100 observations indépendantes de x soit entre 9 et 11.
Si on avait fait 400 observations, la probabilité que $\bar{x}$ soit entre 9 et 11 serait encore plus grande, environ 99,994%, car $\sqrt{400} \cdot \frac{\bar{x}-10}{5}$ serait entre -4 et 4, ce qui correspond à une proportion de 99,994% d'observations de la loi normale standard. De même, la proportion des observations entre 9,5 et 10,5 serait d'environ 68,27% pour $n = 100$ et de 95,45% pour $n = 400$.

V.2.2 TCL appliqué à l'échantillonnage

Ce théorème est *central* pour la science des statistiques (d'où son nom...). Pour nos besoins, il se manifeste de la façon suivante, qui est celle que vous devriez retenir :

Propriété V.2.4

On considère x, une variable quantitative quelconque, suivant une distribution quelconque de moyenne μ et d'écart-type σ. Si le nombre n d'observations est assez grand, la moyenne $\bar{x}$ de n observations indépendantes de x suit une loi normale de moyenne μ et d'écart-type $\sigma/\sqrt{n}$. On peut donc écrire :

$$\bar{x} \sim \mathcal{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right).$$

Que veut dire "n est assez grand" ? Dans presque tous les cas de figure,

$$\boxed{\mathbf{n \geq 30}}$$

est considéré comme suffisant pour que le théorème central limite s'applique. Cependant, si la distribution de x est très asymétrique ou a des queues très épaisses, il peut être nécessaire

d'avoir un n plus grand pour que le théorème central limite s'applique. Inversement, si la distribution de x est "sympathique" : par exemple, unimodale et symétrique, alors un n plus petit peut suffire pour que le théorème central limite s'applique.

Un cas particulier intéressant est celui où x suit déjà une loi normale : dans ce cas, la moyenne $\bar{x}$ de n observations indépendantes de x suit une loi normale de moyenne μ et d'écart-type $\sigma/\sqrt{n}$, quelle que soit la valeur de n. En d'autres termes, dans ce cas, le théorème central limite s'applique même pour un petit nombre d'observations.

ATTENTION : dans cette formulation, c'est la moyenne expérimentale $\bar{x}$ qui suit une loi normale, pas les observations individuelles de x. En effet, les observations individuelles de x suivent la même distribution que x (puisqu'elles sont tirées de cette distribution), qui n'est pas nécessairement une loi normale.

Interprétation du théorème central limite

Le théorème central limite nous rassure sur deux choses que notre intuition nous disait déjà : d'une part, que la moyenne $\bar{x}$ de n observations indépendantes de x est une bonne estimation de la moyenne μ de x, et d'autre part, que plus le nombre n d'observations est grand, plus la moyenne $\bar{x}$ est proche de la moyenne μ (car l'écart-type de $\bar{x}$ est égal à $\sigma/\sqrt{n}$, qui devient de plus en plus petit à mesure que n devient grand). Ce dernier point est une confirmation forte d'un autre principe plus faible : la *loi des grands nombres*, qui dit que la moyenne $\bar{x}$ de n observations indépendantes de x converge vers la moyenne μ de x quand n devient grand, mais sans dire à quelle vitesse cette convergence se fait ni quelle est la distribution de $\bar{x}$ pour un n donné.

Par ailleurs, et c'est un point beaucoup moins intuitif mais très important, le théorème nous dit que le processus d'échantillonnage permet de transformer une distribution de départ quelconque (celle de x), sur laquelle on sait entre rien et pas grand chose, en une distribution de moyenne $\bar{x}$ qui est une loi normale, sur laquelle on sait "tout faire".

V.2.3 Intervalle de confiance

La plus grande force des statistiques est de pouvoir quantifier l'incertitude de nos estimations. Par exemple, si on a un échantillon de 100 observations indépendantes d'une variable x de moyenne μ et d'écart-type σ, on peut calculer la moyenne $\bar{x}$ de ces observations, qui est une estimation de la moyenne μ de x. Cependant, cette estimation que l'on dit *ponctuelle* (car on donne seulement un "point" comme estimé), est incertaine : si on avait tiré un autre échantillon de 100 observations indépendantes de x, on aurait probablement obtenu une moyenne différente. On introduit la notion d'intervalle de confiance pour quantifier cette incertitude.

Définition V.2.5 : Intervalle de confiance

Un intervalle de confiance à p pour cent pour la moyenne μ de x est un intervalle $[a, b]$ (qui dépend de l'échantillon !) tel que la proportion d'observations de $\bar{x}$ entre a et b soit égale à $p\,\%$.
En d'autres termes, si on choisit des échantillons $E_i = \{x_{i,1}, \ldots, x_{i,n}\}$ de n observations indépendantes de x, et qu'on calcule la moyenne $\bar{x}_i$ de chaque échantillon et que pour chaque échantillon on calcule un intervalle de confiance $[a_i, b_i]$ à p pour cent pour la moyenne μ de x, alors la proportion d'intervalles de confiance $[a_i, b_i]$ qui contiennent réellement μ est égale à $p\,\%$.

Propriété V.2.6 : Intervalle de confiance pour une moyenne

Si x suit une loi normale de moyenne μ et d'écart-type σ, alors un intervalle de confiance à 95% pour la moyenne μ de x est donné par la formule suivante :

$$\left[\bar{x} - 1,96 \cdot \frac{\sigma}{\sqrt{n}},\ \bar{x} + 1,96 \cdot \frac{\sigma}{\sqrt{n}}\right].$$

Si on ne connait pas l'écart-type σ de x, on peut le remplacer par l'écart-type s de l'échantillon, et on obtient alors un intervalle de confiance à 95% pour la moyenne μ de x donné par la formule suivante :

$$\left[\bar{x} - 1,96 \cdot \frac{s}{\sqrt{n}},\ \bar{x} + 1,96 \cdot \frac{s}{\sqrt{n}}\right].$$

Remarque V.2.7

Le nombre $1,96$ dans la formule de l'intervalle de confiance à 95% est le nombre a tel que la proportion des observations de la loi normale standard entre $-a$ et a soit égale à 95%, c'est-à-dire telle que :

$$P(-a \leq z \leq a) = 0,95.$$

Si on veut un intervalle de confiance à un autre pourcentage p%, il suffit de remplacer $1,96$ par le nombre d'écarts-types a tel que :

$$P(-a \leq z \leq a) = p\%.$$

Définition V.2.8

On appelle le nombre $1,96 \cdot s/\sqrt{n}$ la *marge d'erreur relative à 95%*. Bien sûr, on peut définir la marge d'erreur relative à $p\%$ en remplaçant $1,96$ par le nombre d'écarts-types a tel que $P(-a \leq z \leq a) = p\%$.

Comme la fonction de répartition de la loi normale est croissante, il est clair que plus on veut un niveau de confiance élevé, plus la taille de l'intervalle de confiance augmente. Inversement, comme la marge d'erreur est inversement proportionnelle à $\sqrt{n}$, plus le nombre n d'observations indépendantes de x est grand, plus la taille de l'intervalle de confiance diminue.

V.2.4 Estimation d'une proportion

Évaluer une proportion à partir d'un échantillon est un cas particulier de l'estimation d'une moyenne : imaginons que la population soit constituée d'individus ayant ou n'ayant pas une caractéristique donnée C (par exemple, des individus qui sont pour ou contre une réforme). On peut alors définir une variable x qui vaut 1 pour les individus ayant la caractéristique C et 0 pour les autres. On peut calculer la moyenne μ_x de x dans la population avec la formule "fréquence absolue", c'est-à-dire :

$$\mu_x = \frac{(1 \times \text{nb d'unités ayant } C) + (0 \times \text{nb d'unités n'ayant pas } C)}{\text{nb total d'unités}} = \frac{\text{nb d'unités ayant } C}{\text{nb total d'unités}},$$

ce qui correspond à la proportion f_C d'individus ayant la caractéristique C dans l'échantillon. Par conséquent, on peut utiliser les outils d'estimation d'une moyenne pour estimer une proportion.

Avant d'utiliser le théorème central limite, il nous faut cependant calculer l'écart-type de cette variable x. Pour fixer les notations, on va noter n_C le nombre d'individus ayant la caractéristique C dans la population de taille N, et donc $n_{\overline{C}} = N - n_C$ le nombre d'individus n'ayant pas la caractéristique C et

$$f_C = \frac{n_C}{N}, \quad f_{\overline{C}} = \frac{n_{\overline{C}}}{N} = \frac{N - n_C}{N} = 1 - f_C,$$

les proportions respectives d'individus ayant ou n'ayant pas la caractéristique C dans la population. On va utiliser la formule "fréquence absolue" de l'écart-type pour une population (en fait pour éviter de faire un calcul avec une grosse racine carrée, on va calculer l'écart-type au

carré, c'est-à-dire la variance) :

$$\begin{aligned}
\sigma^2 &= \frac{(1-\mu_x)^2 \cdot n_C + (0-\mu_x)^2 \cdot n_{\overline{C}}}{N} && \text{On sait que } \mu_x = f_C \text{ donc :} \\
&= (1-f_C)^2 \frac{n_C}{N} + f_C^2 \frac{n_{\overline{C}}}{N} && \text{On remplace } \frac{n_C}{N} \text{ par } f_C \text{ et } \frac{n_{\overline{C}}}{N} \text{ par } 1-f_C : \\
&= (1-f_C)^2 f_C + f_C^2(1-f_C) && \text{On factorise } f_C(1-f_C) : \\
&= f_C(1-f_C)(1-f_C+f_C) && \text{Évidemment, } 1-f_C+f_C = 1 : \\
&= f_C(1-f_C).
\end{aligned}$$

Et donc, en appliquant le théorème central limite, on trouve la propriété suivante, à retenir :

Propriété V.2.9

La proportion mesurée expérimentalement $\hat{p}$ dans un échantillon de taille n assez grande pris dans une population où la proportion réelle est p suit une loi normale :

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) = \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}^2\right).$$

Que veut dire "assez grande" dans ce cas ? En général, on considère que n est assez grand pour que le théorème central limite s'applique si :

$$\boxed{\mathbf{np \geq 5} \text{ et } \mathbf{n(1-p) \geq 5}.}$$

Si on applique le point précédent sur les intervalles de confiance à l'estimation d'une proportion, on obtient :

Propriété V.2.10 : Intervalle de confiance pour une proportion

Si n est assez grand pour que le théorème central limite s'applique, alors un intervalle de confiance à 95% pour la proportion p d'individus ayant la caractéristique C dans la population est donné par la formule suivante :

$$\left[\hat{p} - 1,96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}},\ \hat{p} + 1,96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right].$$

Évidemment, typiquement, on ne connait pas la variance $p(1-p)$, donc on la remplace par l'estimation $\hat{p}(1-\hat{p})$ à partir de l'échantillon, ce qui nous donne la formule précédente. Contrairement à l'estimation d'une moyenne, où en connaissant l'écart-type σ de la variable x, on ne peut pas récupérer la moyenne par un calcul, dans le cas d'une proportion, la situation est tellement simple que si on connait la variance réelle $\sigma^2 = p(1-p)$, on peut directement

en déduire la proportion p : c'est une des deux solutions de l'équation $p(1-p) = \sigma^2$. Cependant, dans la pratique, une solution est au-dessus de 50%, l'autre en dessous, et il s'agit seulement de choisir la bonne. Ainsi, contrairement à l'estimation de la moyenne, il est très rare qu'on connaisse la variance réelle sans connaitre la proportion réelle, et donc on est quasiment toujours obligé de faire l'estimation à partir de l'échantillon.

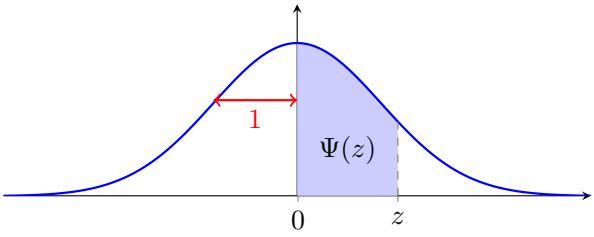
Comme avant, le nombre $1,96$ apparaît dans la formule de l'intervalle de confiance à 95%, mais si on veut un autre niveau de certitude p%, il suffit de remplacer $1,96$ par le nombre d'écarts-types a tel que :

$$P(-a \leq z \leq a) = p\%.$$

Digression 1 : la table du livre.

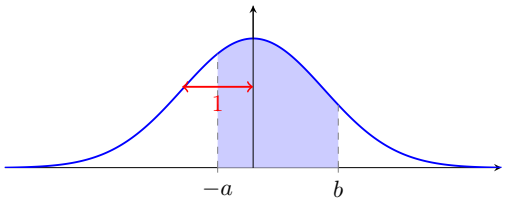
Vous verrez parfois[1], une table de la loi normale standard qui donne la proportion d'observations entre 0 et une valeur positive de z :

$$\Psi(z) = \int_0^z \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx =$$

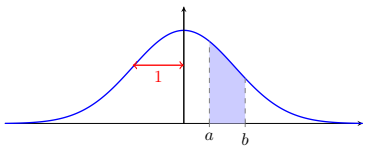


L'utilisation de Ψ pour calculer l'aire sous la courbe en est légèrement différente :

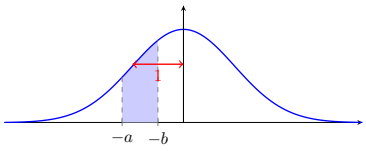
$\Psi(a) + \Psi(b) =$



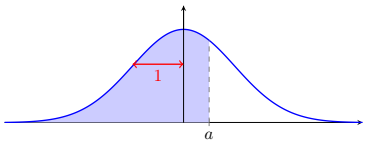
$\Psi(b) - \Psi(a) =$



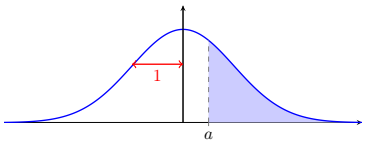
$\Psi(a) - \Psi(b) =$



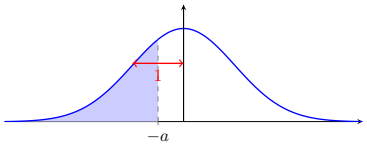
$0,5 + \Psi(a) =$



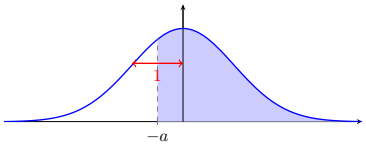
$0,5 - \Psi(a) =$



$0,5 - \Psi(a) =$



$0,5 + \Psi(a) =$



Bien entendu, si on a affaire à une loi normale non standard, on se ramène à ce qui peut être lu dans cette table via la cote z.

1. En particulier dans votre manuel page 224.

Digression 2 : estimateur sans biais

Définition V.2.11 : Estimateur

Un estimateur $\hat{\theta}$ d'un paramètre θ est une statistique, calculée à partir d'un échantillon, qui est utilisée pour estimer la valeur du paramètre θ dans la population.

Exemple V.2.12

La moyenne $\bar{x}$ d'un échantillon de taille n est un estimateur de la moyenne μ de la population. L'écart-type expérimental s d'un échantillon de taille n est un estimateur de l'écart-type σ de la population.

Un estimateur peut être bon ou mauvais : par exemple, si on prend la statistique "maximum de l'échantillon" pour estimer le paramètre "moyenne de la population", on s'attend (à juste titre) à ce que cet estimateur soit très mauvais. Moins trivialement, il est tout à fait possible que la formule qui donne le paramètre quand on l'applique à la population ne soit pas la même que la formule qui donne le meilleur estimateur à partir de l'échantillon : par exemple, la formule de l'écart-type d'une population est différente de la formule de l'écart-type d'un échantillon, et pourtant, c'est la formule de l'écart-type d'un échantillon qui est un meilleur estimateur de l'écart-type de la population que la formule de l'écart-type d'une population appliquée à l'échantillon. Pour quantifier la qualité d'un estimateur, on peut utiliser la notion de biais :

Définition V.2.13 : Biais d'un estimateur

Le biais d'un estimateur $\hat{\theta}$ d'un paramètre θ est défini comme la différence entre la moyenne de l'estimateur sur tous les échantillons possibles et la valeur réelle du paramètre. Un estimateur est dit sans biais si son biais est égal à zéro, c'est-à-dire si, en moyenne, il donne la bonne valeur du paramètre.

Propriété V.2.14

La moyenne $\bar{x}$ d'un échantillon de taille n est un estimateur sans biais de la moyenne μ de la population.

Donnons nous quelques notations pour pouvoir en faire la preuve :

— On note E un échantillon de taille n pris dans la population de taille N.

— On note N_E le nombre d'échantillons différents de taille n que l'on peut former à partir de la population de taille N.

— On note $\overline{x}_E$ la moyenne de l'échantillon E.

Démonstration. On veut vérifier si

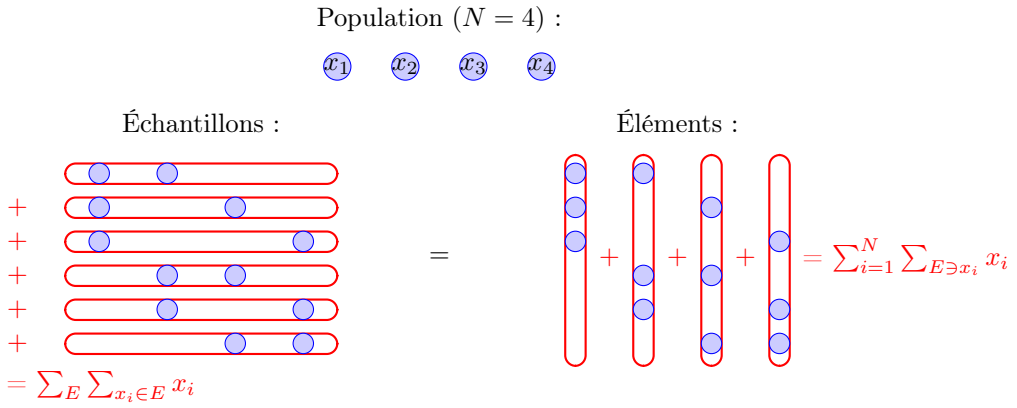
$$\frac{1}{N_E}\sum_E \overline{x}_E = \mu.$$

Remplaçons $\overline{x}_E$ par sa formule :

$$\begin{aligned}\frac{1}{N_E}\sum_E \overline{x}_E &= \frac{1}{N_E}\sum_E \frac{1}{n}\sum_{x_i\in E} x_i \\ &= \frac{1}{nN_E}\sum_{i=1}^{N}\sum_{E\ni x_i} x_i.\end{aligned}$$

Entre la première et la deuxième ligne, on a interverti les deux sommes : au lieu de faire la somme sur les échantillons E puis sur les éléments x_i de chaque échantillon, on fait d'abord la somme sur les éléments x_i de la population, puis pour chaque élément x_i, on fait la somme sur les échantillons E qui contiennent cet élément x_i. De plus, comme n ne dépend ni de E ni de x_i, on peut le sortir de la somme.

FIGURE V.7 – Changement de sommation : de la somme sur les échantillons à la somme sur les éléments de la population.



Maintenant, remarquons que la somme interne est simplement $x_i + ... + x_i$, autant de fois que le nombre d'échantillons E qui contiennent l'élément x_i. Or, ce nombre ne dépend en fait pas de i : x_i n'est pas spécial et il y a donc autant d'échantillons qui contiennent x_i que d'échantillons qui contiennent x_j pour tout j. Par conséquent, on peut remplacer la somme interne par $x_i \cdot k$, où k est le nombre d'échantillons qui contiennent un élément donné de la population.

$$\frac{1}{N_E}\sum_E \overline{x}_E = \frac{1}{nN_E}\sum_{i=1}^{N} kx_i = \frac{k}{nN_E}\sum_{i=1}^{N} x_i.$$

Pour conclure, il faut que $\frac{k}{nN_E} = \frac{1}{N}$, ce qui implique que $kN = nN_E$. Or :

$$\begin{aligned} kN = &\text{ nombre d'échantillons contenant } x_1 \\ &+ ... \\ &+ \text{ nombre d'échantillons contenant } x_N \end{aligned}$$

Dans cette somme, chacun des N_E échantillons de taille n est compté exactement n fois, car il contient exactement n éléments de la population (par exemple, si $n = 3, N = 7$ et un échantillon contient x_2, x_3, x_7, il apparait dans deuxième, troisième et septième termes de la somme). Par conséquent, la somme totale est égale à nN_E, ce qui conclut la preuve. □

Une propriété similaire est vraie pour la variance.

Propriété V.2.15

La variance expérimentale $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ d'un échantillon de taille n est un estimateur sans biais de la variance σ^2 de la population.

La démonstration est considérablement plus technique si on ne dispose pas d'outils supplémentaires.

Résumé du chapitre

Loi normale et courbe de Gauss

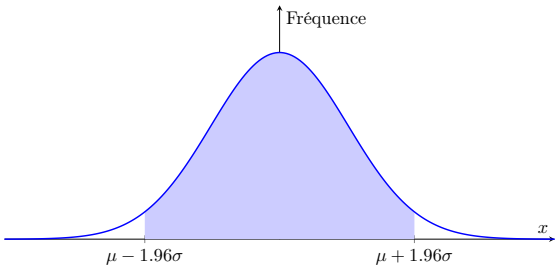


FIGURE V.8 – 95% des observations se situent dans $[\mu - 1,96\sigma, \mu + 1,96\sigma]$

Propriétés clés :

- Symétrique autour de la moyenne μ
- Largeur contrôlée par l'écart-type σ
- Très commune dans la nature (tailles, poids, QI, etc.)
- Permet de calculer des probabilités

Théorème central limite (TCL) Si on prélève n observations indépendantes d'une variable quelconque (même pas normale !) et qu'on les moyenne, la moyenne suit une loi normale.

Cas	Condition	Résultat	Implication
Moyenne	$n \geq 30$	$\bar{x} \sim \mathcal{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$	Plus n est grand, plus $\bar{x}$ converge vers μ
Proportion	$np \geq 5$ et $n(1-p) \geq 5$	$\hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}^2\right)$	Plus n est grand, plus $\hat{p}$ converge vers p

Intervalle de confiance Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors l'intervalle $[\mu - 1,96\sigma, \mu + 1,96\sigma]$ contient 95% des observations de X. Selon le cas (moyenne ou proportion) et ce qu'on connait (paramètres ou statistiques), on remplace μ et σ par les formules correspondantes, ce qui nous donne des formules d'intervalle de confiance à 95% pour μ ou p.

En pratique :

$$\mu \in \left[\bar{x} - 1,96 \cdot \frac{s}{\sqrt{n}}, \quad \bar{x} + 1,96 \cdot \frac{s}{\sqrt{n}}\right] \quad \text{et} \quad p \in \left[\hat{p} - 1,96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \quad \hat{p} + 1,96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right].$$

$$\boxed{\text{avec 95\% de confiance}} \text{ (il } \textit{faut} \text{ le dire !)}$$

Interprétation :

- 95% de confiance que cet intervalle contienne μ ou p (selon le cas).
- À confiance fixée, plus n augmente, plus l'intervalle rétrécit. À n fixé, plus la confiance augmente, plus l'intervalle s'élargit.
- Si on veut un niveau de confiance de $99\%, 99,9\%, 99,99\%, 99,999\%$ etc, à la place de 1,96 on utilise respectivement 2,58, 3,29, 3,89, 4,42 etc.

Appendice : les valeurs de loi normale (à partir de 0).

z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1.0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1.1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1.2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1.3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1.4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1.5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1.6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1.7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1.8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1.9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2.0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2.1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2.2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2.3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2.4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2.5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2.6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2.7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2.8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2.9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3.0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3.1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3.2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3.3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3.4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3.5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3.6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3.7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3.8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3.9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Chapitre VI

Analyse bivariée

Dans ce chapitre, nous allons décrire comment *deux* variables changent ensemble.

— Quels types de lien peut-il y avoir entre deux variables ?
— Comment détecter et mesurer ce lien ?
— Comment utiliser ce lien pour faire des prédictions ?

VI.1 Types de lien entre deux variables

Rappelons les notions de variable indépendante et variable dépendante, vues dans le chapitre 2.

Définition VI.1.0 : Variable indépendante et variable dépendante

Dans une expérience, la variable **indépendante** est celle que le chercheur manipule, tandis que la variable **dépendante** est celle qui est mesurée pour observer la réponse à la manipulation.

On rappelle que contrairement au langage courant, la variable dépendante ne dépend pas nécessairement de la variable indépendante : cette terminologie reflète le rôle joué par les variables dans une expérience, et non pas la nature de leur relation. Notons aussi que la "manipulation" de la variable indépendante peut être plus ou moins artificielle : dans une expérience, le chercheur peut faire varier la variable indépendante en assignant les participants à différents groupes, ou en leur faisant suivre différentes procédures. Mais dans une étude observationnelle, le chercheur peut simplement mesurer la variable indépendante telle qu'elle existe naturellement, sans intervenir et utiliser ses variations naturelles pour expliquer les variations de la variable dépendante.

VI.1.1 Causalité

Définition VI.1.1 : Relation de causalité

Dire que la variable X cause la variable Y signifie que la présence ou l'amplitude de X est directement responsable de la présence ou de l'amplitude de Y. En d'autres termes, la variable Y dépend de la variable X au sens habituel du terme "dépendance" : Y ne peut pas exister ou ne peut pas prendre certaines valeurs sans X. On dit que X est la variable causale de Y.

Dans une expérience étudiant une relation de causalité, la variable indépendante est la cause et la variable dépendante est l'effet. La relation est à sens unique et faire artificiellement varier la variable d'effet ne fera pas varier la variable causale.

Exemple VI.1.2

Le tabagisme est une cause du cancer du poumon : fumer des cigarettes augmente le risque de développer un cancer du poumon, mais développer un cancer du poumon ne fait pas augmenter le risque de devenir fumeur.

VI.1.2 Influence (mutuelle)

Dans la réalité, il est relativement rare que les liens de cause et d'effet entre deux variables soient clairs et surtout "totaux" au sens où une variable explique complètement l'autre. Souvent, pour un effet donné, les causes sont multifactorielles et il n'est pas rare que des rétroactions fassent que la variable "d'effet" entraine des variations dans la variable de "cause". Dans ces cas où il existe une influence mais où la causalité n'est pas aussi nette que pour un lien de causalité à proprement parler, on parle, sans surprise de *lien d'influence*.

Définition VI.1.3 : Relation d'influence

Deux variables X et Y sont dites en relation d'influence si elles varient ensemble de manière que les variations de X soient responsables des variations de Y, et/ou que les variations de Y soient responsables des variations de X.

Exemple VI.1.4

L'inflation est, grossièrement, causée par le fait que "trop" d'argent se dispute "trop peu" de biens. Une manière de faire baisser l'inflation est d'encourager les gens à garder leur argent dans leur épargne, en augmentant les taux d'intérêts. Ainsi, il y a un lien entre les taux d'intérêts et l'inflation, et ce lien est bidirectionnel. En effet, trop d'inflation est mauvaise pour l'économie et une inflation élevée pousse les banques centrales à augmenter leur taux directeur qui, à son tour, pousse l'inflation vers le bas.

VI.1.3 Concomitance

On l'a déjà vu dans la première partie du cours, des grandeurs peuvent varier ensemble sans que l'une soit la cause de l'autre, ni même que les deux s'influencent mutuellement : par exemple, les parapluies et les bottes de pluie sont concomitants, car ils sont tous les deux liés à la pluie, mais ils ne causent pas l'un l'autre et ne s'influencent pas mutuellement.

Définition VI.1.5 : Relation de concomitance

Deux variables X et Y sont dites concomitantes si elles varient ensemble sans que l'une soit la cause de l'autre, ni que les deux s'influencent mutuellement.

Le cas le plus simple expliquant la concomitance entre deux variables est celui où les deux variables sont causées par une troisième variable.

Exemple VI.1.6

Le nombre de visites à l'opéra, en concert, etc, et la valeur de la voiture possédée sont positivement liés, mais aucun des deux ne cause l'autre : les deux sont influencés positivement par le revenu disponible.

VI.1.4 Indépendance

À l'opposé de la causalité, il peut y avoir indépendance entre deux variables : les variations de l'une n'ont aucune influence sur les variations de l'autre.

Définition VI.1.7

Deux variables X et Y sont dites indépendantes si les variations de X n'ont aucune influence sur les variations de Y, et inversement. En d'autres termes, la distribution de Y est la même pour toutes les valeurs de X, et la distribution de X est la même pour toutes les valeurs de Y.

Exemple VI.1.8

Le résultat d'un lancer de dé à six faces est indépendant du résultat d'un autre lancer de dé à six faces : les variations de l'un n'ont aucune influence sur les variations de l'autre et la distribution des résultats est la même pour les deux lancers. Inversement, le fait d'être allé au cinéma et le fait d'avoir mangé du pop-corn ne sont pas indépendants, car parmi les personnes allant au cinéma, la proportion de celles ayant mangé du pop-corn est plus élevée que dans la population générale.

VI.2 Test d'hypothèse : le χ^2

VI.2.1 Généralités sur les tests statistiques

L'idée générale d'un test statistique pour différencier entre deux hypothèses H_0 et H_1 est de construire un nombre à partir d'un échantillon de données, appelé *statistique de test*, qui va prendre des valeurs "extrêmes" si l'hypothèse nulle H_0 est fausse et des valeurs "modérées" si l'hypothèse nulle H_0 est vraie. On calcule cette statistique sur notre échantillon, puis on se pose la question de savoir à quel point l'observation d'une valeur au moins aussi extrême que celle que l'on a observée est probable si H_0 est vraie. Si l'observation est très improbable (à un niveau fixé à l'avance que l'on appelle le *seuil de signification*) sous H_0, alors on rejette H_0 au profit de H_1. Inversement, si l'observation donne quelque chose "d'attendu" sous H_0, on continue de croire que H_0 est vraie (au moins provisoirement) et on ne rejette pas H_0 au profit de H_1.

Exemple VI.2.0

Un professeur soupçonne un groupe d'étudiants d'avoir triché lors d'un examen. Il constate que les notes des étudiants sont normalement distribuées dans tout le groupe. Il décide de calculer la moyenne et l'écart-type des notes de la classe (disons $\mu = 64$ et $\sigma = 20$), puis de calculer la moyenne des 15 élèves soupçonnés de tricherie. Il trouve une moyenne de $\bar{x} = 76$. Si le professeur se trompe, et que les étudiants accusés forment en fait un échantillon aléatoire de la classe, le théorème central limite nous dit que la probabilité d'observer une moyenne de 76 ou plus est $P(\bar{x} \geq 76) = P(Z \geq (76-64)/(20/\sqrt{15})) = P(Z > 2,32) \approx 0,01$ (où Z suit une loi normale centrée réduite). Le professeur conclut que les étudiants soupçonnés de tricherie ne forment pas un échantillon aléatoire de la classe et qu'au moins une partie d'entre eux a triché.

Il existe une grande variété de tests statistiques, chacun ayant des caractéristiques différentes et étant adapté à des situations différentes.

Définition VI.2.1

La *confiance* d'un test statistique est la probabilité de ne pas rejeter l'hypothèse nulle H_0 lorsqu'elle est vraie. Elle est égale à $1-\alpha$, où α est le *niveau de signification* du test, c'est-à-dire la probabilité de rejeter H_0 lorsqu'elle est vraie (erreur de type I).
La *puissance* d'un test statistique est la probabilité de rejeter l'hypothèse nulle H_0 lorsqu'elle est fausse. Elle est égale à $1-\beta$, où β est la probabilité de ne pas rejeter H_0 lorsqu'elle est fausse (erreur de type II).

TABLE VI.1 – Issues possibles d'un test statistique.

Décision du test	Réalité	
	H_0 est vraie	H_0 est fausse
H_0 n'est pas rejetée	Vrai négatif ($1-\alpha$ confiance du test)	Faux négatif (erreur de type II)
H_0 est rejetée	Faux positif (erreur de type I)	Vrai positif ($1-\beta$ puissance du test)

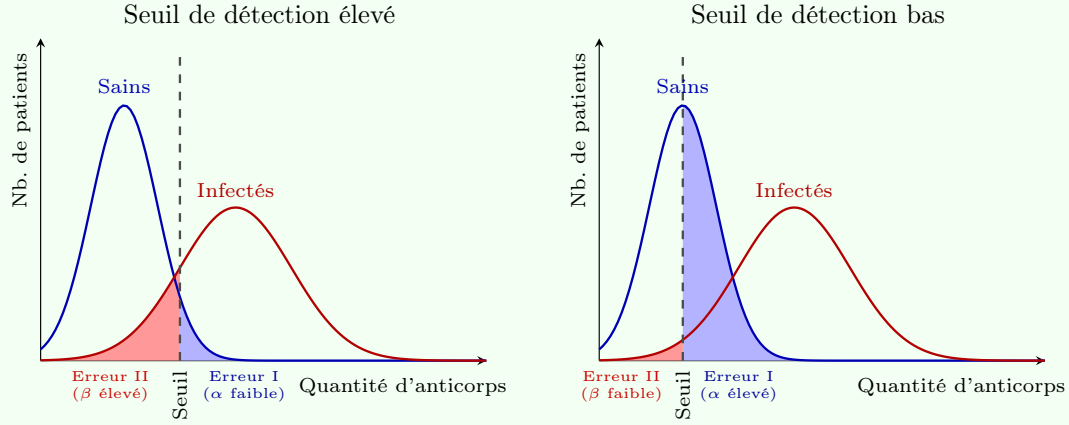
Attention : la confiance et la puissance ne permettent pas, en elles-mêmes, de calculer la probabilité de prendre une bonne décision à partir d'un test statistique. La confiance et la puissance sont des propriétés du test lui-même, tandis que la probabilité de prendre une bonne décision dépend aussi de la réalité des échantillons/populations/unités statistiques étudiées.

Exemple VI.2.2

Dans cet exemple, on utilise le cas d'un test médical, pas d'un test statistique en général, mais les idées présentées sont les mêmes pour tous les tests.

Imaginons que l'on veuille tester la présence d'une maladie chez un patient. La maladie est repérée à partir du niveau d'anticorps dans le sang du patient : si le niveau d'anticorps est supérieur à un certain seuil, alors on considère que le test est positif ("la maladie est détectée"), sinon on considère qu'il est négatif ("la maladie n'est pas détectée").

Imaginons de plus que l'on connaisse la répartition du niveau d'anticorps chez les patients sains et chez les patients infectés.



Dans le cas d'un test médical, il est souvent préférable d'avoir un seuil de détection bas, afin de minimiser les faux négatifs (erreur de type II), même si cela signifie avoir un taux de faux positifs plus élevé (erreur de type I), pour être sûr de détecter les patients ayant besoin de traitement. On préférera donc le test de droite, dans lequel il est peu

probable de ne pas détecter un patient infecté (grande puissance), même si cela signifie que certains patients sains seront détectés à tort comme infectés (faible confiance).
Si on ajoute à cela l'information que seule une personne sur 10 000 est infectée, si une personne fait un test qui revient positif, le risque que cette personne soit malade est en fait assez faible, car il y a beaucoup plus de chance (a priori) que cette personne soit sur la courbe bleue, où elle a de grande chance d'être détectée par erreur. C'est pour cela que dans le cas de maladies graves avec des traitements lourds, on préfère souvent faire un test de dépistage avec un seuil de détection bas, pour être sûr de détecter les patients ayant besoin de traitement, puis faire un test de confirmation avec un seuil de détection plus élevé, pour être sûr de ne pas traiter des patients sains.

VI.2.2 Test du χ^2

Imaginons que l'on choisisse 100 personnes au hasard au Québec, et qu'on leur fasse chacun tirer une pièce, puis qu'on enregistre le genre de la personne et le résultat du tirage. Comme il y a autant d'hommes que de femmes au Québec (à très peu de choses près), on s'attend à ce qu'on ait 50% d'hommes et, si la pièce n'est pas truquée, que 50% d'entre eux tirent pile, pour un total de 25% d'hommes qui tirent pile. De même, on s'attend à ce que 25% d'hommes tirent face, 25% de femmes tirent pile et 25% de femmes tirent face. Imaginons que l'on observe la répartition suivante :

TABLE VI.2 – Une expérience imaginaire.

Genre	Résultat du tirage		Total
	Pile	Face	
Homme	31	19	50
Femme	18	32	50
Total	49	51	100

Il est à peu près clair que les hommes ont tiré plus de piles que les femmes, mais d'un autre côté, c'est un processus aléatoire, il serait étonnant que les valeurs observées soient exactement égales aux valeurs attendues. Est-ce que cette différence est suffisamment grande pour conclure que le genre et le résultat du tirage ne sont pas indépendants ? Par le théorème central limite, les valeurs observées se comportent comme des variables normales. L'idée, pour mesurer si la différence observée est grande au point d'être "surprenante", est de considérer la distance entre les valeurs observées et les valeurs attendues, en la normalisant par l'écart-type de la distribution des valeurs observées (ainsi, si une fréquence relative attendue de 0,0001 donne 0,01, on compte cela comme plus surprenant qu'une fréquence relative attendue de 0,1

donnant 0,2). Par le TCL, c'est une somme de carrés de lois normales.

Définition VI.2.3

La loi du χ^2 à k degrés de liberté décrit la distribution des sommes des carrés de k variables aléatoires indépendantes suivant une loi normale centrée réduite. En d'autres termes, si $Z_1, Z_2, \ldots, Z_k$ sont k variables aléatoires indépendantes suivant une loi normale centrée réduite, alors la variable aléatoire $X = Z_1^2 + Z_2^2 + \ldots + Z_k^2$ suit une loi du χ^2 à k degrés de liberté.

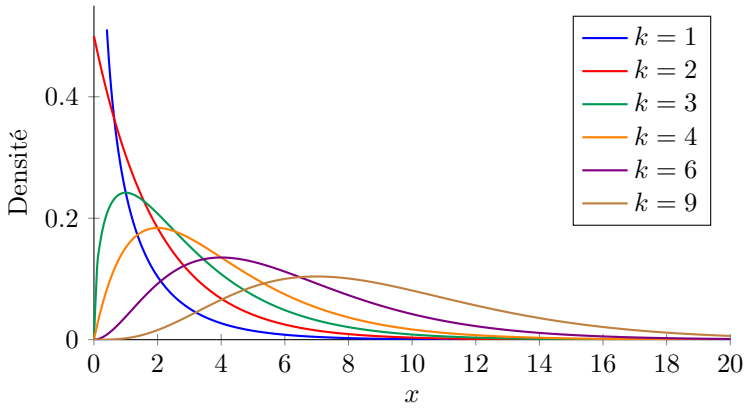


FIGURE VI.1 – Fonctions de densité du χ^2 pour différents degrés de liberté.

Définition VI.2.4 : Statistique du χ^2

Soient deux variables X et Y ayant respectivement r et c modalités. On note n le nombre total d'observations, O_{ij} le nombre d'observations ayant la modalité i pour la variable X et la modalité j pour la variable Y. On note également R_i le nombre d'observations ayant la modalité i pour la variable X, et C_j le nombre d'observations ayant la modalité j pour la variable Y. La statistique du χ^2 est définie par la formule suivante :

$$\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

où $E_{ij} = \frac{R_i C_j}{n}$ est le nombre d'observations attendu dans la cellule ij sous l'hypothèse d'indépendance entre les variables X et Y.

L'idée de la statistique du χ^2 est de comparer les données observées à ce que l'on attendrait d'observer si les deux variables étaient indépendantes. On s'attend, même si les variables sont indépendantes, à une petite déviation entre les données observées et les données attendues : la statistique χ^2_{obs} mesure l'ampleur de cette déviation. Par le théorème central limite, O_{ij} suit

une loi normale dont la moyenne est, si X et Y sont indépendantes, E_{ij}. Sous l'hypothèse d'indépendance, la statistique T suit donc une loi du χ^2 [1]. Inversement, si les variables ne sont pas indépendantes, une des quantités $O_{ij} - E_{ij}$ sera significativement différente de zéro, ce qui fera que la statistique T sera significativement plus grande que ce que l'on attendrait d'observer si les variables étaient indépendantes : si la probabilité d'observer une statistique T aussi grande que celle que l'on a observée, ou plus grande, est inférieure à un seuil de signification (généralement 5%), alors on rejette l'hypothèse d'indépendance entre les variables X et Y.

Exemple VI.2.5

Dans notre exemple de pièce et de genre, la statistique du χ^2 est égale à :

$$\chi^2_{obs} = \frac{(31-24.5)^2}{24.5} + \frac{(19-25.5)^2}{25.5} + \frac{(18-24.5)^2}{24.5} + \frac{(32-25.5)^2}{25.5} = 6,56.$$

Propriété VI.2.6

Si les variables X et Y sont indépendantes, alors la statistique du χ^2 suit une loi du χ^2 à $(r-1)(c-1)$ degrés de liberté.

Pourquoi $(r-1)(c-1)$ degrés de liberté ? Il y a r modalités pour la variable X et c modalités pour la variable Y, ce qui fait rc cellules dans le tableau de contingence. Cependant, les totaux de chaque ligne et de chaque colonne sont fixés par les données observées, ce qui fait que seulement $(r-1)(c-1)$ cellules sont libres de varier : une fois que les valeurs de $(r-1)(c-1)$ cellules sont fixées, les valeurs des autres cellules sont déterminées par les totaux de chaque ligne et de chaque colonne. Comme la loi du χ^2 est définie à partir de la somme des carrés de variables aléatoires *indépendantes*, le nombre de degrés de liberté correspond au nombre de cellules libres de varier, soit $(r-1)(c-1)$ d'entre elles.

Exemple VI.2.7

Dans notre exemple de pièce et de genre, il y a $r = 2$ modalités pour la variable "genre" (homme et femme) et $c = 2$ modalités pour la variable "résultat du tirage" (pile et face), ce qui fait que la statistique du χ^2 suit une loi du χ^2 à $(2-1)(2-1) = 1$ degré de liberté.

Dans l'exemple suivant, on a une variable nominale (tendance politique) avec 3 modalités (droite, centre, gauche) et une variable de rapport avec 4 classes ($[15-25[$, $[25-45[$, $[45-65[$, > 65), ce qui fait que la statistique du χ^2 suit une loi du χ^2 à $(4-1)(3-1) = 6$

1. Les variables $O_{ij} - E_{ij}$ sont clairement centrées, mais il faut un peu plus de travail mathématique pour montrer que diviser par E_{ij} suffit à les réduire dans leur ensemble.

degrés de liberté.

TABLE VI.3 – Répartition d'une population selon la classe d'âge et la tendance politique.

Classe d'âge	**Tendance politique**			**Total**
	Droite	**Centre**	**Gauche**	
$[15, 25[$	14	13	29	56
$[25, 45[$	15	20	19	54
$[45, 65[$	22	19	10	51
> 65	9	11	3	23
Total	60	63	61	184

Il est important de vérifier que les conditions d'application du test du χ^2 sont respectées : généralement, il faut que tous les effectifs attendus E_{ij} soient supérieurs à 5 et que $n \geq 30$ pour que le test soit valide. Cette condition vient essentiellement du fait que l'on approxime le contenu de chaque case du tableau croisé par une loi normale, et que cette approximation est plus fiable lorsque les effectifs attendus sont suffisamment grands.

Exemple VI.2.8

Dans notre exemple de pièce et de genre, c'est bien le cas, car toutes les fréquences théoriques sont supérieures à 24.

Enfin, on se pose la question de savoir si la valeur observée est "surprenante". Pour cela, il faut d'abord choisir ce que veut dire "être surpris".

Définition VI.2.9

Le seuil de signification α d'un test statistique est la probabilité de rejeter l'hypothèse nulle H_0 lorsqu'elle est vraie (erreur de type I). En d'autres termes, c'est le niveau de risque que l'on est prêt à accepter pour rejeter H_0 à tort. Un seuil de signification de 5% signifie que l'on accepte un risque de 5% de rejeter H_0 alors qu'elle est en réalité vraie.

Dans le cas du test du χ^2, qui teste l'existence d'un lien entre variables, le seuil de signification est ce que l'on considère être le risque acceptable de conclure à tort que les variables sont liées alors qu'elles sont en réalité indépendantes. On veut donc savoir si observer χ^2_{obs} (ou plus) est improbable ou non si on suppose que les variables sont indépendantes.

Définition VI.2.10 : p-valeur

La p-valeur d'un test statistique est la probabilité, sous l'hypothèse nulle H_0, d'observer une statistique de test au moins aussi extrême que celle qui a été observée. Dans le cas du test du χ^2, c'est la valeur de $P(T \geq \chi^2_{obs})$ où T suit une loi du χ^2 à $(r-1)(c-1)$ degrés de liberté.

Exemple VI.2.11

Dans notre exemple de pièce et de genre, avec $\chi^2_{obs} = 6,56$ et $(r-1)(c-1) = 1$ degré de liberté, la p-valeur est égale à $P(T \geq 6,56) \approx 0,01$.
Comme on le voit sur l'exemple suivant, la p-valeur dépend à la fois de la valeur de χ^2_{obs} et du nombre de degrés de liberté. Par exemple, si on avait observé $\chi^2_{obs} = 6,56$ avec 9 degrés de liberté, la p-valeur aurait été beaucoup plus grande, soit $P(T \geq 6,56) \approx 0,68$.

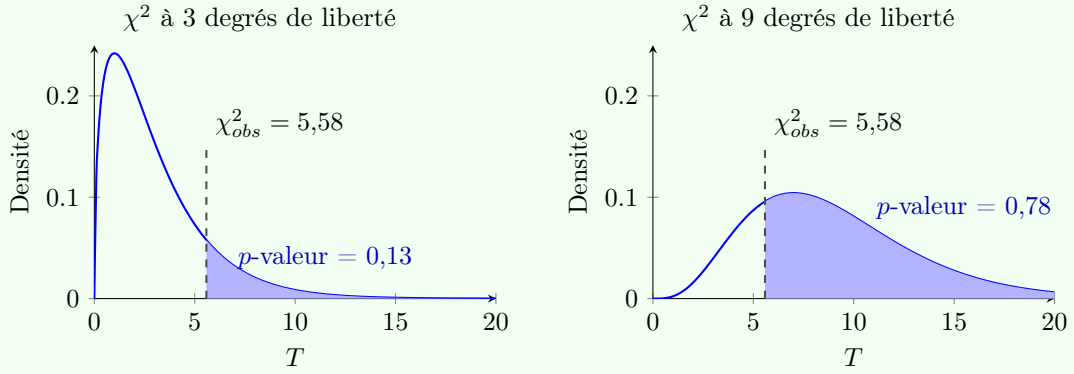


FIGURE VI.2 – Comparaison des p-valeurs pour deux tests du χ^2 avec différents degrés de liberté.

Si notre p-valeur est inférieure à notre seuil de signification α, on considère le résultat "trop surprenant" sous l'hypothèse H_0 et on rejette H_0 au profit de H_1. Inversement, si notre p-valeur est supérieure à notre seuil de signification α, on considère le résultat "pas si surprenant" sous l'hypothèse H_0 et on ne rejette pas H_0 au profit de H_1.

Malheureusement, il est difficile de calculer la p-valeur à la main étant donné χ^2_{obs}. On peut remarquer que la p-valeur est décroissante en fonction de χ^2_{obs} : plus χ^2_{obs} est grand, plus la p-valeur est petite. On peut donc se contenter de comparer χ^2_{obs} à la valeur critique de la loi du χ^2 à $(r-1)(c-1)$ degrés de liberté pour le seuil de signification α, c'est-à-dire la valeur exacte telle que $P(T \geq \text{valeur critique}) = \alpha$.

Exemple VI.2.12

Dans notre exemple de pièce et de genre, avec $\alpha = 0,05$ et $(r-1)(c-1) = 1$ degré de liberté, la valeur critique est égale à $\chi^2_{critique} = 3,84$.

Si χ^2_{obs} est supérieur à la valeur critique, alors la p-valeur est inférieure à α et on rejette H_0. Inversement, si χ^2_{obs} est inférieur à la valeur critique, alors la p-valeur est supérieure à α et on ne rejette pas H_0.

Exemple VI.2.13

Dans notre exemple, comme $\chi^2_{obs} = 6,56$ est supérieur à la valeur critique $\chi^2_{critique} = 3,84$, on rejette l'hypothèse d'indépendance entre le genre et le résultat du tirage.

Tout cela se résume par la méthodologie suivante pour réaliser un test du χ^2 d'indépendance entre deux variables qualitatives (ou quantitatives regroupées en classes) X et Y :

Méthode VI.2.14

— Préparation du test.

1. Choix ou identification des variables X et Y à tester.
2. Recensement du nombre de modalités ou classes de chaque variable : l pour la variable X et c pour la variable Y.
3. Formulation des hypothèses : H_0 = "les variables X et Y sont indépendantes" et H_1 = "les variables X et Y ne sont pas indépendantes".
4. Choix du seuil de signification α (généralement 5%).
5. Récolte ou récupération des données.

— Réalisation du test.

1. On présente les données sous la forme d'un tableau croisé des fréquences absolues observées O_{ij}, où i correspond à la modalité i de la variable X et j correspond à la modalité j de la variable Y.
2. On calcule les totaux de chaque ligne L_i et de chaque colonne C_j, ainsi que le total général (qui est la taille de l'échantillon) n.
3. Pour chaque paire de modalités i et j, on calcule le nombre d'observations attendu $E_{ij} = \frac{L_i C_j}{n}$ sous l'hypothèse d'indépendance.
4. On vérifie que les conditions d'application du test du χ^2 sont respectées : généralement, il est recommandé que tous les effectifs attendus E_{ij} soient supérieurs à 5 pour que le test soit valide.

5. On calcule la statistique du χ^2 : $\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$.
6. On compare la statistique χ^2_{obs} à la valeur critique de la loi du χ^2 à $(l-1)(c-1)$ degrés de liberté pour le seuil de signification α.
7. Si χ^2_{obs} est supérieur à la valeur critique, on rejette l'hypothèse nulle H_0 et on conclut que les variables X et Y ne sont pas indépendantes. Sinon, on ne rejette pas l'hypothèse nulle H_0 et on conclut que les données ne fournissent pas suffisamment de preuves pour rejeter l'indépendance entre les variables X et Y.

TABLE VI.4 – Organisation du tableau croisé et notations pour le test du χ^2.

Variable X	Variable Y					Total lignes
	Y_1	$\cdots$	Y_j	$\cdots$	Y_c	
X_1	O_{11}	$\cdots$	O_{1j}	$\cdots$	O_{1c}	L_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\iddots$	$\vdots$	$\vdots$
X_i	O_{i1}	$\cdots$	O_{ij}	$\cdots$	O_{ic}	L_i
$\vdots$	$\vdots$	$\iddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_l	O_{l1}	$\cdots$	O_{lj}	$\cdots$	O_{lc}	L_l
Total colonnes	C_1	$\cdots$	C_j	$\cdots$	C_c	n

Interprétation du test du χ^2 L'interprétation standard du résultat du test du χ^2 est selon le cas :

"On accepte (ou on rejette) l'hypothèse H_1 au seuil de signification α."

ou, de façon équivalente :

"On rejette (ou on n'a pas suffisamment de preuves pour rejeter) l'hypothèse H_0 au seuil de signification α."

Typiquement, pour être plus clair, on préfèrera remplacer H_0 et H_1 par leur signification concrète. Par exemple, dans notre exemple de pièce et de genre, on dira plutôt :

"On rejette l'hypothèse d'indépendance entre le genre et le résultat du tirage au seuil de signification de 5%."

Coefficient de Cramér Comme les valeurs attendues du χ^2 dépendent du nombre de degrés de liberté, il est difficile de comparer les résultats de différents tests du χ^2 entre eux. Le coefficient de Cramér est une mesure d'association entre deux variables qualitatives qui permet

de comparer les résultats de différents tests du χ^2 entre eux, indépendamment du nombre de degrés de liberté. De la même façon, quand la taille de l'échantillon augmente, même si le théorème central limite garantit que l'écart relatif entre les effectifs observés et les effectifs théoriques devient très petit ($\sim 1/\sqrt{n}$), l'écart absolu peut devenir très grand ($\sim \sqrt{n}$), ce qui fait que la statistique du χ^2 peut devenir très grande même si les variables sont presque indépendantes. Le coefficient de Cramér permet également de corriger ce problème en normalisant la statistique du χ^2 par la taille de l'échantillon.

Définition VI.2.15 : Coefficient de Cramér

Le coefficient de Cramér, noté V, est défini par la formule suivante :

$$V = \sqrt{\frac{\chi^2_{obs}}{n \cdot (\min(l, c) - 1)}}$$

où χ^2_{obs} est la statistique du χ^2 observée, n est le nombre total d'observations, l est le nombre de modalités de la variable X et c est le nombre de modalités de la variable Y.

Plus le coefficient de Cramér est proche de 0, plus les variables sont indépendantes. Plus le coefficient de Cramér est proche de 1, plus les variables sont associées. On peut l'interpréter de la manière suivante :

— $V = 0$: les variables sont indépendantes.
— $0 < V < 0,1$: il existe une association très faible entre les variables.
— $0,1 \leq V < 0,2$: il existe une association faible entre les variables.
— $0,2 \leq V < 0,3$: il existe une association moyenne entre les variables.
— $0,3 \leq V < 0,4$: il existe une association forte entre les variables.
— $V \geq 0,4$: il existe une association très forte entre les variables.

Ces niveaux d'interprétations sont conventionnels et peuvent varier en fonction du contexte de l'étude. Il est important de noter que le coefficient de Cramér ne mesure que la force de l'association entre les variables, et non la direction de cette association.

VI.3 Corrélation et régression linéaire

Dans le cas de deux variables quantitatives, on a à notre disposition davantage d'outils mathématiques. En premier lieu, comme chaque donnée peut être placée sur une droite numérique, on peut représenter les données par un nuage de points dans le plan, pour visualiser la relation entre les deux variables.

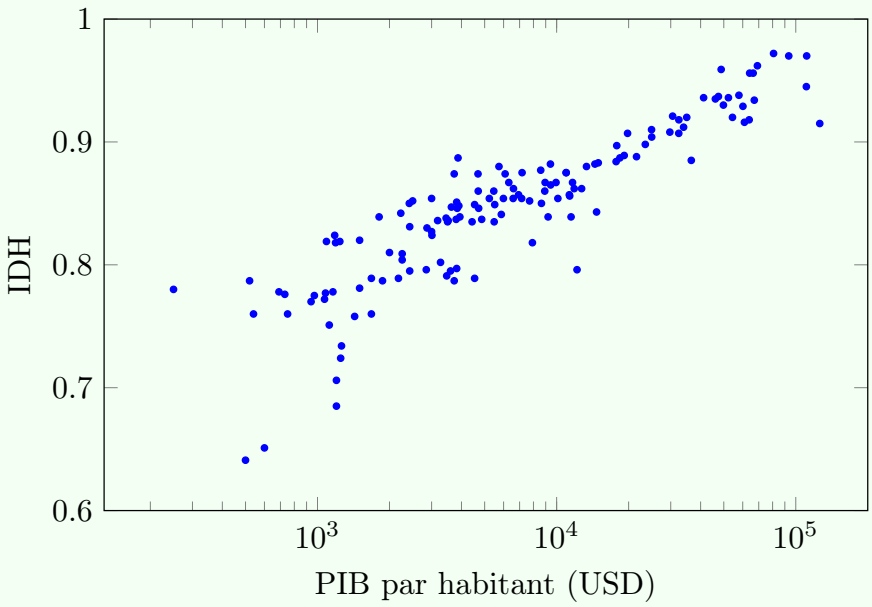
Définition VI.3.0 : Nuage de points

Le nuage de points d'une série statistique à deux variables quantitatives X et Y est l'ensemble des points du plan de coordonnées (x_i, y_i), où x_i et y_i sont les valeurs de X et Y respectivement pour l'observation i, pour $i = 1, 2, \ldots, n$.

Exemple VI.3.1

Considérons un ensemble de données contenant le PIB par habitant (en USD) et l'Indice de Développement Humain (IDH) de 2023 pour plusieurs pays (Source : Banque mondiale, Programme des Nations Unies pour le développement). Nous voulons visualiser la relation entre ces deux variables quantitatives en créant un nuage de points.

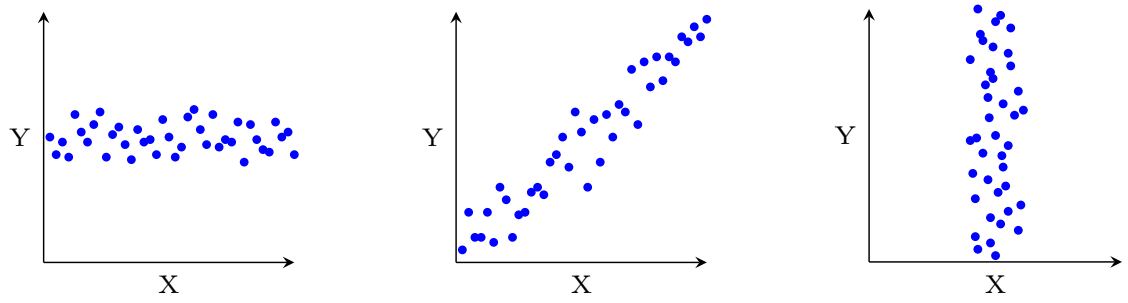
FIGURE VI.3 – Répartition de 141 pays selon leur PIB par habitant et leur IDH.



Il apparait clairement qu'il existe une relation entre PIB par habitant et IDH : quand le PIB/hab augmente, l'IDH tend à augmenter également. Ce n'est pas très surprenant, étant donné que le PIB par habitant est l'un des indicateurs utilisés pour calculer l'IDH. Il faut cependant remarquer que l'IDH n'est pas entièrement déterminé par le PIB par habitant, autrement il n'y aurait aucune variation de l'IDH pour un même PIB par habitant, ce qui n'est pas le cas : on observe une certaine dispersion des points du nuage de points pour un même PIB par habitant, ce qui suggère que d'autres facteurs que le PIB par habitant influencent également l'IDH. Notons également que l'axe horizontal n'est pas linéaire, mais logarithmique, ce qui indique que la relation entre les deux

grandeurs n'est pas si simple qu'elle peut paraitre au premier coup d'œil.

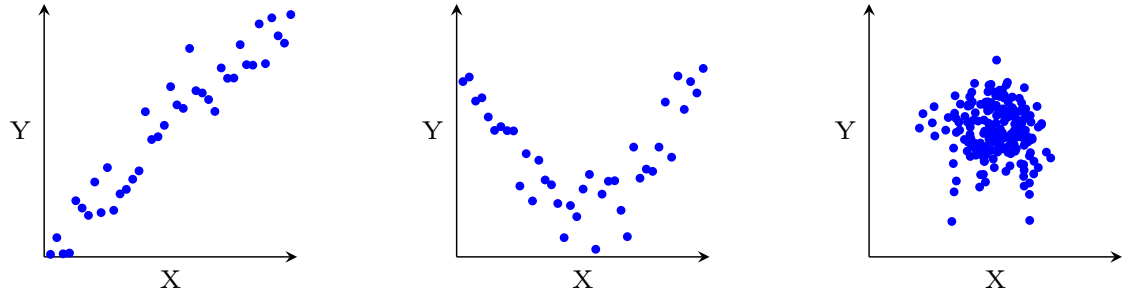
D'une façon générale, une relation entre X et Y apparait dans le nuage de points comme une "forme" autour de laquelle les points sont regroupés, et de façon importante, qui varie à la fois en fonction de X et de Y. Par exemple, dans la figure ci-dessous, X et Y ne sont liés que dans le nuage (b), où étant donné la valeur de X, les valeurs de Y sont contraintes à être proches d'une certaine valeur qui dépend de X. Au contraire, dans les nuages (a) et (c), X et Y ne sont pas liés. Dans le nuage (a), varier X n'a aucune influence sur les valeurs de Y, qui sont réparties de manière homogène quelle que soit la valeur de X et inversement dans le nuage (c).



(a) Y ne dépend pas de X. (b) X et Y sont liés. (c) X ne dépend pas de Y.

FIGURE VI.4 – Un alignement des points ne signifie pas forcément qu'il y a une relation.

Plusieurs types de relations peuvent exister entre X et Y.



(a) Relation linéaire. (b) Relation non linéaire. (c) Pas de relation

FIGURE VI.5 – Différents types de relations entre X et Y.

VI.3.1 Corrélation

L'*existence* de toutes sortes de relations peut être établie via le test du χ^2, mais on peut également tenter de décrire certaines relations en plus grands détail en utilisant les opérations

arithmétiques offertes par les variables quantitatives. En particulier, en ce qui concerne le type de relation le plus simple qui soit, une relation "linéaire", où le nuage de points se répartit autour d'une droite, on peut mesurer à quel point le nuage de points est bien approximé par une droite à l'aide du *coefficient de corrélation linéaire*.

Définition VI.3.2

Le coefficient de corrélation linéaire entre deux variables quantitatives X et Y est défini par la formule suivante :

$$r = \frac{1}{(n-1)s_X s_Y} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

où n est le nombre d'observations, s_X et s_Y sont les écarts-types de X et Y respectivement, et $\bar{x}$ et $\bar{y}$ sont les moyennes de X et Y respectivement. On peut donc le réécrire comme :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Si on a affaire à une population au lieu d'un échantillon, on peut remplacer $n-1$ par n, $\bar{x}, \bar{y}$ par les moyennes théoriques μ_X, μ_Y et s_X, s_Y par les écarts-types théoriques σ_X, σ_Y dans la formule.

Propriété VI.3.3

Le coefficient de corrélation linéaire r est compris entre -1 et 1.

Plus r est proche de 1, plus les points du nuage de points sont alignés selon une droite de pente positive. Plus r est proche de -1, plus les points du nuage de points sont alignés selon une droite de pente négative. Si r est égal à 0, cela signifie qu'il n'y a pas de corrélation linéaire entre les variables X et Y, c'est-à-dire que les points du nuage de points ne sont pas alignés selon une droite[2]. La force de la corrélation linéaire peut être interprétée comme indiqué à gauche.

TABLE VI.5 – Interprétation de la force de la corrélation.

Valeur de $\|r\|$	**Interprétation**
$0 \leq \|r\| < 0{,}1$	Nulle
$0{,}10 \leq \|r\| < 0{,}25$	Très faible
$0{,}25 \leq \|r\| < 0{,}50$	Faible
$0{,}50 \leq \|r\| < 0{,}75$	Modérée
$0{,}75 \leq \|r\| < 0{,}90$	Forte
$0{,}90 \leq \|r\| \leq 1$	Très forte à parfaite

2. Ou, si c'est le cas, que la droite est parallèle à un des axes et qu'on ne peut donc rien déduire sur une variable à partir de l'autre.

VI.3.2 Régression

Une fois déterminé que X et Y sont liés par une relation linéaire, on peut tenter de trouver la droite qui "s'ajuste" le mieux au nuage de points, c'est-à-dire la droite qui minimise la distance entre les points du nuage de points et la droite.

Définition VI.3.4

La droite de régression linéaire de Y sur X est la droite d'équation $y = a + bx$ qui approxime le mieux le nuage de points (au sens de minimiser la somme des carrés des distances verticales entre les points du nuage de points et la droite).

Propriété VI.3.5

Les coefficients a et b de la droite de régression linéaire de Y sur X peuvent être calculés à partir des moyennes et des écarts-types des variables X et Y, ainsi que du coefficient de corrélation linéaire r : $\quad b = r\dfrac{s_Y}{s_X}, \quad a = \bar{y} - b\bar{x}.$

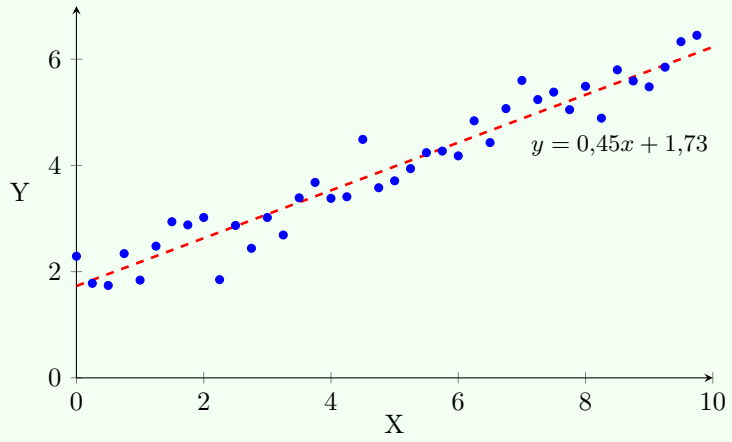
Bien sûr, si on considère une population au lieu d'un échantillon, on peut remplacer les moyennes et les écarts-types par leurs versions théoriques dans la formule ci-dessus.

Définition VI.3.6

Dans le cas d'une série temporelle, la droite de régression linéaire de Y sur X est appelée *droite de tendance* de la série temporelle.

Exemple VI.3.7

FIGURE VI.6 – Exemple de droite de régression linéaire : $y = 0,45x + 1,73$.



Ici, le coefficient de corrélation est $r = 0{,}97$, la moyenne de X est $\bar{x} = 4{,}88$, l'écart-type de X est $s_X = 2{,}92$, la moyenne de Y est $\bar{y} = 3{,}95$ et l'écart-type de Y est $s_Y = 1{,}37$, ce qui donne bien $b = 0,97 \times 1{,}37/2{,}92 = 0{,}45$ et $a = 3{,}95 - 0{,}45 \times 4{,}88 = 1{,}73$.

Enfin, on peut se demander à quel point la variable indépendante X explique la variable dépendante Y. En d'autres termes, on peut se demander quelle proportion de la variance de Y est expliquée par la relation linéaire entre X et Y.

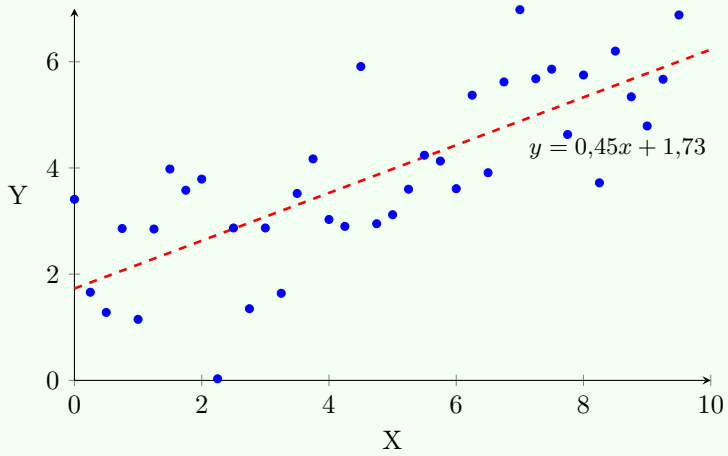
Définition VI.3.8

Le *coefficient de détermination* est défini comme r^2, le carré du coefficient de corrélation linéaire. C'est la proportion de la variance de Y qui est expliquée par la relation linéaire entre X et Y.

Exemple VI.3.9

Dans l'exemple précédent, le coefficient de détermination est $r^2 = 0{,}93$, ce qui signifie que 93% de la variance de Y est expliquée par la relation linéaire entre X et Y.
Si on reprend le même nuage de points mais qu'on l'étire verticalement, on augmente la variance de Y sans changer la relation linéaire entre X et Y, ce qui fait que le coefficient de corrélation linéaire diminue, et donc que le coefficient de détermination diminue également. Ainsi, dans le diagramme suivant, qui a la même droite de régression, le coefficient de détermination est de $r^2 = 0{,}60$

FIGURE VI.7 – Exemple avec un coefficient de détermination de $r^2 = 0{,}60$.



Dans ce nouvel exemple, le coefficient de corrélation linéaire est $r = 0{,}78$, ce qui signifie que la relation linéaire entre X et Y est plus faible que dans l'exemple précédent. Le

coefficient de détermination est alors $r^2 = 0{,}60$, ce qui indique que seulement 60% de la variance de Y est expliquée par la relation linéaire avec X.

On peut également se demander si la relation linéaire entre X et Y est statistiquement significative, c'est-à-dire si elle est suffisamment forte pour ne pas être due au hasard : si on regarde trois points pris au hasard, il y a de bonnes chances pour que l'on puisse faire passer une droite près des trois. Au contraire, si un nuage de 1000 points est très bien ajusté par une droite, ce n'est probablement pas dû au hasard de la mesure. Ainsi, plus la taille de l'échantillon est faible, plus il faut que la relation soit forte pour être considérée comme significative.

TABLE VI.6 – Valeurs critiques du coefficient de corrélation linéaire pour différents niveaux de signification et différentes tailles d'échantillon.

Seuil de signification	**Taille de l'échantillon**					
	5	10	20	50	100	1000
5%	0,878	0,632	0,444	0,279	0,197	0,062
2%	0,934	0,715	0,516	0,328	0,232	0,074
1%	0,959	0,765	0,561	0,361	0,256	0,081

Interprétation des statistiques de régression On vérifie d'abord que le coefficient de corrélation linéaire r est supérieur à la valeur critique pour le seuil de signification choisi et la taille de l'échantillon. Si c'est le cas, on considère que la relation linéaire entre X et Y est statistiquement significative, et on peut commenter sur la force et la direction de cette relation, ainsi que sur la proportion de la variance de Y qui est expliquée par cette relation linéaire (coefficient de détermination r^2).

Par exemple, si on a $n = 10, r = -0,78, r^2 = 0,60$ et $\alpha = 5\%$ on l'interprète comme :

"Il existe une relation linéaire négative forte entre X et Y, qui est statistiquement significative au seuil de 5%. Environ 60% de la variance de Y est expliquée par cette relation linéaire avec X."

Inversement, si on a $n = 10, r = 0{,}50, r^2 = 0{,}25$ et $\alpha = 5\%$, on l'interprète comme :

"Il n'existe pas de relation linéaire statistiquement significative au seuil de 5%.

On peut ensuite donner une interprétation plus concrète à partir de l'équation de la droite de régression linéaire. Supposons qu'on ait établi qu'il existe une relation significative entre X et Y et qu'on ait calculé la droite de régression linéaire de Y sur X donnée par l'équation

$y = ax + b$. On peut alors interpréter le coefficient de pente a comme la variation moyenne de Y pour une unité de variation de X.

"En moyenne, pour chaque augmentation de 1 unité de X, Y augmente de a unités."

Souvent, le coefficient a n'est pas entier, par exemple $a = 1,25 = 5/4$. Dans ce cas, pour rendre l'interprétation plus concrète, on peut dire "En moyenne, pour chaque augmentation de 4 unités de X, Y augmente de 5 unités".

VI.3.3 Prédictions ?

Si on a réussi à extraire une droite de régression ou de tendance des données, on peut l'utiliser pour inférer des valeurs de Y à partir de valeurs de X : si on a une droite de régression linéaire de Y sur X donnée par l'équation $y = a + bx$, alors pour une valeur donnée de X, disons x_0, on peut prédire la valeur correspondante de Y en calculant $y_0 = a + bx_0$.

Exemple VI.3.10

Dans notre exemple précédent, si on veut prédire la valeur de Y pour $X = 6$, on peut calculer $y = 0{,}45 \times 6 + 1{,}73 = 4{,}43$.

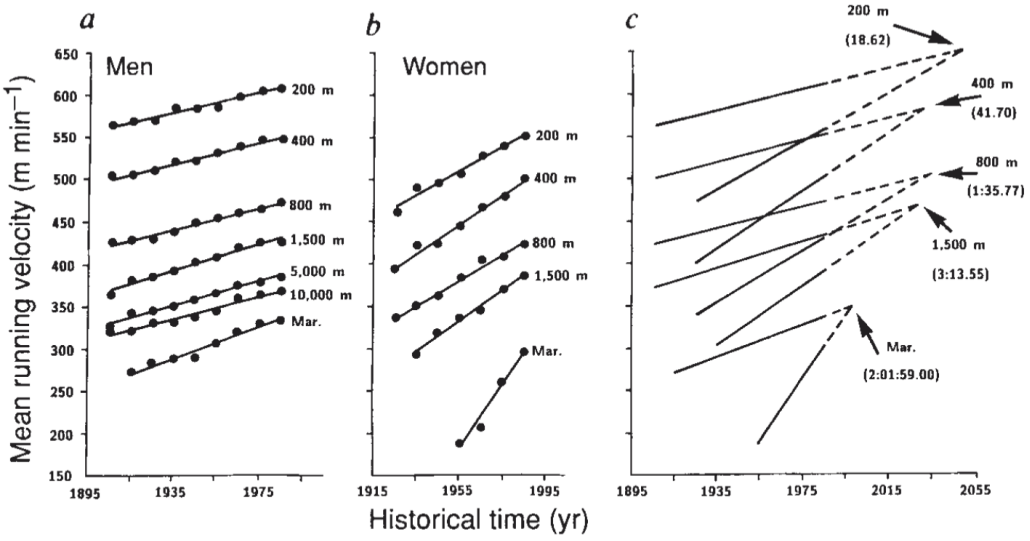
Cependant, il est important de noter que ces prédictions ne sont fiables que si la relation linéaire entre X et Y est suffisamment forte (c'est-à-dire si le coefficient de corrélation linéaire est élevé) et si les données utilisées pour construire la droite de régression sont représentatives de la population ou du phénomène que l'on souhaite modéliser.

Il n'est pas raisonnable d'utiliser une estimation linéaire trop loin des données. Par exemple, si on a effectué nos mesures sur une population dont le patrimoine est inférieur à 1 million de dollars, les estimations linéaires pour des patrimoines de l'ordre de 10 milliards de dollars ont toutes les chances d'être très mauvaises. Il est courant d'obtenir une bonne approximation linéaire sur une partie des données sans que la relation globale soit linéaire. De même, lorsqu'il s'agit d'une série temporelle, il est risqué d'utiliser une droite de tendance pour faire des prédictions dans un futur lointain, car la tendance peut changer au fil du temps.

Exemple VI.3.11

Un exemple célèbre d'une mauvaise utilisation d'une droite de tendance est donné dans la correspondance *Will women soon outrun men ?* par Brian J. Whipp & Susana Ward, publiée dans le journal *Nature* en 1992. Dans cette note, les auteurs font une régression linéaire sur les temps de marathon des hommes et de femmes et prédisent que les femmes pourraient battre les temps masculins en 1998. En 2026, le record féminin est de 2h09 :56, environ 10 minutes de plus que le record masculin, et 9 minutes de plus

que le temps prédit par l'étude pour 1998.



World record progression, expressed as mean running velocity versus historical time, for men (*a*) and women (*b*), with best-fit linear regressions (solid lines) superimposed. In *c*, the regression lines for the common events for men and women (solid lines) are extrapolated (dashed lines) to their points of intersection; the predicted world record times at these intersection points are shown in parentheses (h:min:s)

Les auteurs n'ont pas pris en compte que les gains rapides des femmes dans les années 1980 étaient principalement dus à l'augmentation du nombre de femmes participant à des marathons, et que ce gain ne pouvait pas se poursuivre indéfiniment.

L'absurdité du modèle est claire si on le pousse plus loin encore : assez loin dans le futur, le modèle prédit que les femmes courraient le marathon en un temps négatif, ce qui est évidemment impossible.