Nº 12-004-X au catalogue

Les statistiques : le pouvoir des données!

Date de diffusion : le 2 septembre 2021

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Table des matières

Les statistiques : le pouvoir des données!

Les statistiques : le pouvoir des données! est un outil de formation destiné principalement aux étudiants, mais également aux enseignants et à tous ceux qui désirent tirer pleinement parti des statistiques. Il vise à aider le lecteur à :

- apprendre à utiliser avec plus d'assurance l'information statistique,
- reconnaître l'importance de l'information statistique dans la société d'aujourd'hui,
- apprendre à utiliser avec un esprit critique les données qui lui sont présentées.

Ces objectifs sont au cœur de la mission de Statistique Canada d'appuyer les Canadiens dans la prise de décisions éclairées et fondées sur les données.

La première section de cette ressource définit les concepts de données, de statistiques et d'information statistique, ainsi que de qualité des données. La seconde section décrit les types de données pouvant être utilisées pour produire l'information statistique. La troisième section détaille les différentes étapes du processus de production d'information statistique, telles que l'échantillonnage, la collecte de données, la vérification et l'imputation, l'estimation et le couplage d'enregistrements. Les sections 4 et 5 expliquent comment utiliser les statistiques descriptives et la visualisation des données pour explorer les données. Chaque chapitre a été conçu de manière à être complet en lui-même et contient des exercices qui ont pour but d'aider à mieux assimiler la matière présentée.

Il est important de noter que la longueur et le niveau de détails d'une section ne reflètent aucunement l'importance du sujet dans le processus global de production d'information statistique, mais plutôt le niveau technique attendu du public visé par la ressource. Par exemple, bien que la visualisation des données soit un outil pertinent à l'école secondaire, les détails de l'estimation des données requièrent des connaissances avancées dans le domaine de la statistique mathématique qui sont généralement acquises au cours d'études postsecondaires.

Il faut finalement souligner l'influence de la deuxième édition de la publication **Statistics — A Powerful Edge!** du Bureau de la statistique de l'Australie qui a servi de base à l'élaboration de cette publication électronique canadienne au début des années 2000 et dont l'influence est tout aussi grande dans cette version récemment mise à jour de **Les statistiques : le pouvoir des données !**

1 Les données, l'information statistique et les statistiques

Notre société est de plus en plus riche en données et elle fait face à des défis nouveaux lorsque vient le temps de filtrer et d'interpréter cette masse d'information toujours grandissante. Plus que jamais auparavant, les administrations publiques, l'industrie et les citoyens ont besoin d'informations statistiques fiables pour prendre de meilleures décisions, mais la tâche de fournir de l'information de grande qualité en temps opportun ne cesse de croître en complexité.

Ce besoin d'informer la société est l'une des raisons pour lesquelles le système d'éducation canadien développe un curriculum axé sur la collecte, le traitement et la présentation des données. Avant qu'un utilisateur d'information puisse se lancer dans de telles activités, il est toutefois important qu'il comprenne bien les termes **données**, **information statistique** et **statistiques**, et qu'il se donne un moyen d'en évaluer la qualité.

1.1 Définitions

Les données, l'information statistique et les statistiques sont étroitement liées. Pour naviguer à travers l'océan de plus en plus vaste des informations produites par la société moderne, il est important de comprendre les différences clés entre ces trois concepts. Les données sont les matériaux bruts pour la production d'information statistique et les statistiques sont un type particulier d'information statistique.

Données

Les données sont des faits, des chiffres, des observations ou des enregistrements qui peuvent se présenter sous la forme d'image, de son, de texte ou de mesure physique (p. ex. distance, poids, longueur d'onde). Les données peuvent être collectées et traitées dans le but de tirer des conclusions. Les données peuvent provenir de plusieurs sources et elles peuvent être divisées en deux groupes en fonction de la forme qu'elles prennent : les données structurées et les données non structurées.

Les données structurées sont organisées en éléments prédéfinis, chacun correspondant à un concept ou à un élément d'information spécifique. Un ensemble de données collectées en utilisant un questionnaire ou un formulaire à remplir est un bon exemple de données structurées : les questions représentent des concepts séparés et bien définis. Dans le cas d'une question fermée, la réponse se trouvera parmi l'une des catégories prédéfinies pour ce concept. Dans le cas d'une question ouverte, la réponse pourrait prendre la forme d'une valeur numérique ou d'un texte. Si une valeur a été obtenue pour chacun des concepts ou éléments d'information, les données sont complètes. Sinon, elles contiennent des valeurs manquantes.

Regardons par exemple la manière dont chaque colonne du tableau 1.1.1 sur les universités canadiennes est en lien avec un concept distinct :

Tableau 1.1.1
Exemple de données structurées

Nom de l'établissement	Ville	Province	Date de fondation	Nombre d'étudiants
Université Laval	Québec	QC	1852	43 000
Université de Waterloo	Waterloo	ON	1955	30 000
Université Dalhousie	Halifax	NE	1818	18 000
Université Simon Fraser	Burnaby	CB	1965	30 000

Chaque rangée présente les valeurs des variables d'une unité d'observation pour laquelle l'information a été recueillie. Les rangées sont désignées comme des observations ou des enregistrements. Les concepts présentés dans chaque colonne sont souvent appelés des variables. Les ensembles de données sont des regroupements de données qui ont les mêmes définitions pour les unités d'observation et les variables.

Pour être traitées et analysées, les données structurées doivent être compilées dans une structure de données digitale qui s'aligne sur les concepts prédéfinis ou les variables, telles qu'une feuille de calcul, une base de données ou un fichier texte délimité. Les données peuvent ensuite être importées dans un logiciel statistique qui permet à l'utilisateur des données de les transformer, de les agréger, de procéder à des opérations mathématiques sur les données ou de les visualiser.

Les données non structurées sont n'importe quelles données qui ne sont pas organisées selon un modèle prédéfini. Pour produire de l'information statistique à partir des données non structurées, un traitement additionnel des données est nécessaire pour organiser l'information. Le tableau 1.1.2 présente la façon dont un texte, une image ou un enregistrement vocal peuvent être convertis en données structurées pour l'analyse textuelle, la reconnaissance des images et la reconnaissance du langage.

Tableau 1.1.2
Transformer des données non structurées en données structurées

Données non structurées	Traitement	Données structurées
Un texte	Découpage du texte en une liste de mots; agrégation pour compter le nombre d'occurrences de chaque mot; utilisation de dictionnaires et de règles pour classer les mots	Une feuille de calcul : chaque rangée correspond à un mot distinct, trois colonnes présentent le mot, la fréquence du mot dans le texte et la catégorie du mot
Une image	Attribution d'un code RVB à chaque pixel; segmentation de l'image en groupes de pixels en fonction des composantes rouges (R), vertes (V) et bleues (B).	Une base de données : chaque enregistrement correspond à un groupe de pixels et les champs résument les composantes de couleur de chaque groupe.
L'enregistrement de la voix d'une personne	Segmentation de l'enregistrement en sons distincts; mesure des durées et fréquences de chaque son.	Une liste des segments accompagnés de leur durée et de leur fréquence.

Avec l'utilisation accrue des ordinateurs et des téléphones intelligents dans tous les domaines de la vie quotidienne, une partie énorme de l'information qui est créée aujourd'hui est non structurée. Évaluer le potentiel de ces données et trouver des façons innovantes de les rassembler, les traiter et les analyser pour produire de l'information statistique de valeur est l'un des grands défis de la révolution des données.

Mais quelle est la différence entre l'information statistique et les données?

Information statistique

L'information statistique est constituée de données qui ont été enregistrées, classées, organisées, reliées ou interprétées à l'intérieur d'un cadre conceptuel de façon à ce qu'un sens en émerge. L'information statistique qui est présentée aux utilisateurs de l'information doit les aider à comprendre l'histoire que les données racontent, mais également leur communiquer la qualité de l'information présentée. Elle peut être présentée dans des formats variés : textes, tableaux, graphiques, infographies, vidéos ou même bases de données.

Plusieurs exemples d'information statistique produite à Statistique Canada seront présentés à la page suivante, mais il est d'abord important de comprendre une étape incontournable du processus de production l'information statistique : l'utilisation des statistiques!

Statistiques

Les statistiques sont généralement en lien avec les données numériques. Le terme **statistique** peut faire référence à la discipline scientifique qui s'intéresse à l'analyse des données numériques. Les statistiques, quant à elles, sont un type d'information obtenu en soumettant les valeurs à des opérations mathématiques. Avant tout, l'objectif des statistiques est de fournir une information utile aux utilisateurs par le moyen des nombres.

Les types de statistiques les plus communément utilisées pour résumer l'information statistique sont appelés les statistiques descriptives. Pour les variables numériques, les mesures de tendance centrale correspondent à la valeur la plus représentative des unités trouvées dans un ensemble de données. Les mesures de dispersion correspondent à l'étalement des valeurs autour de la tendance centrale. Pour les variables catégoriques, les distributions de fréquences sont utilisées pour résumer les données. Les proportions, les ratios et les taux sont également des statistiques descriptives utiles pour l'analyse des données.

Lorsqu'un ensemble de données contient sur chaque ligne des statistiques qui résument l'information de plusieurs unités d'observation, il s'agit d'un ensemble de données agrégées. À l'opposé, lorsque l'ensemble de données contient l'information d'une seule unité d'observation sur chaque ligne, il s'agit d'un ensemble de microdonnées.

1.2 Exemples d'information statistique

Comme vous allez voir, l'information statistique peut être présentée de plusieurs façons, telles que les textes, les tableaux, les visualisations ou les infographies.

Enquête trimestrielle sur les marchandises vendues au détail

Ceci est un exemple d'information statistique qui peut être utilisée pour la prise de décision. À Statistique Canada, la diffusion de nouvelles informations statistiques sur les indicateurs économiques clés est souvent faite par le biais d'une courte communication écrite dans Le Quotidien.

> Au deuxième trimestre de 2019, les ventes au détail au Canada ont atteint 163,3 milliards de dollars, en hausse de 1,4 % par rapport au même trimestre en 2018. Les ventes ont augmenté dans 13 des 19 groupes de marchandises au deuxième trimestre de 2019.
>
> Les ventes d'aliments ont enregistré la plus forte hausse en dollars, lesquelles se sont accrues de 3,5 % d'une année à l'autre. La majeure partie de la hausse était attribuable à une augmentation des ventes d'aliments frais (+3,4 %), plus particulièrement des ventes de fruits et de légumes frais (+5,4 %). Les ventes d'aliments secs emballés ont augmenté de 4,4 %, alors que les ventes d'aliments congelés ont progressé de 1,1 %. Les ventes de boissons gazeuses et de boissons alcoolisées ont enregistré une hausse de 2,3 % comparativement à l'année précédente, en grande partie en raison d'une augmentation des ventes des boissons alcoolisées (+2,6 %).
>
> Les ventes de pièces, d'accessoires et de fournitures pour véhicules automobiles ont crû de 5,0 % au deuxième trimestre. Cette hausse était en grande partie attribuable aux ventes accrues de pièces et d'accessoires de véhicules automobiles (+6,0 %) et de pneus neufs de véhicules automobiles (+3,6 %).
>
> Parallèlement, les ventes de véhicules automobiles ont augmenté de 0,6 % au cours de la même période en raison d'une hausse des ventes de véhicules automobiles d'occasion (+6,7 %). Les recettes tirées de la vente de véhicules automobiles neufs ont diminué de 2,5 %. Cette baisse était principalement attribuable à une diminution des ventes de voitures particulières neuves (-14,9 %).
>
> Les ventes d'articles de quincaillerie, d'outils, d'articles de rénovation et de produits pour pelouse et jardin ont augmenté de 1,9 % au deuxième trimestre. La hausse des ventes d'équipement et d'articles domestiques pour la pelouse et le jardin (+8,2 %) a contribué le plus à cette augmentation.
>
> Les ventes de carburants pour les véhicules automobiles et de combustibles résidentiels ont diminué de 2,7 % par rapport au deuxième trimestre de 2018. Ce recul était en grande partie attribuable à une diminution des ventes de carburants pour les véhicules automobiles (-2,6 %).
>
> Les ventes de produits du cannabis se sont chiffrées à 251,8 millions de dollars au deuxième trimestre de 2019. Les ventes de sommités fleuries de cannabis séchées ont atteint 222,4 millions de dollars, tandis que les ventes d'huile de cannabis se sont chiffrées à 29,0 millions de dollars.
>
> Source : Statistique Canada, Le Quotidien, 15/10/2019

Professions, Recensement de 1911

Voici un tableau d'information statistique portant sur les professions au Canada au cours du dernier siècle. On peut y voir le nombre de Canadiens qui exerçaient des professions particulières au moment du Recensement de 1911. Il convient de noter que certaines professions avaient une désignation différente à cette époque!

Tableau 1.2.1
Certaines professions tirées du Recensement du Canada de 1911

Professions	Hommes	Femmes
Gardiens de pont et de porte	436	2
Hommes et femmes de ménage	12	4 700
Blanchisseurs et blanchisseuses	588	282
Gardiens d'hôtel	3 102	848
Entrepreneurs de pompes funèbres	43	0
Jardiniers	469	18
Cochers et palefreniers	418	0
Matelots et marins	16 347	0
Marieurs	72	178
Infirmiers	124	5 476
Sténographes et dactylos	1 603	9 754
Acteurs et gens de théâtre	2 410	432
Musiciens et professeurs de musique	2 001	3 574
Colporteurs et marchands de rues	3 135	113
Total	**30 760**	**25 377**

0 zéro absolu ou valeur arrondie à zéro

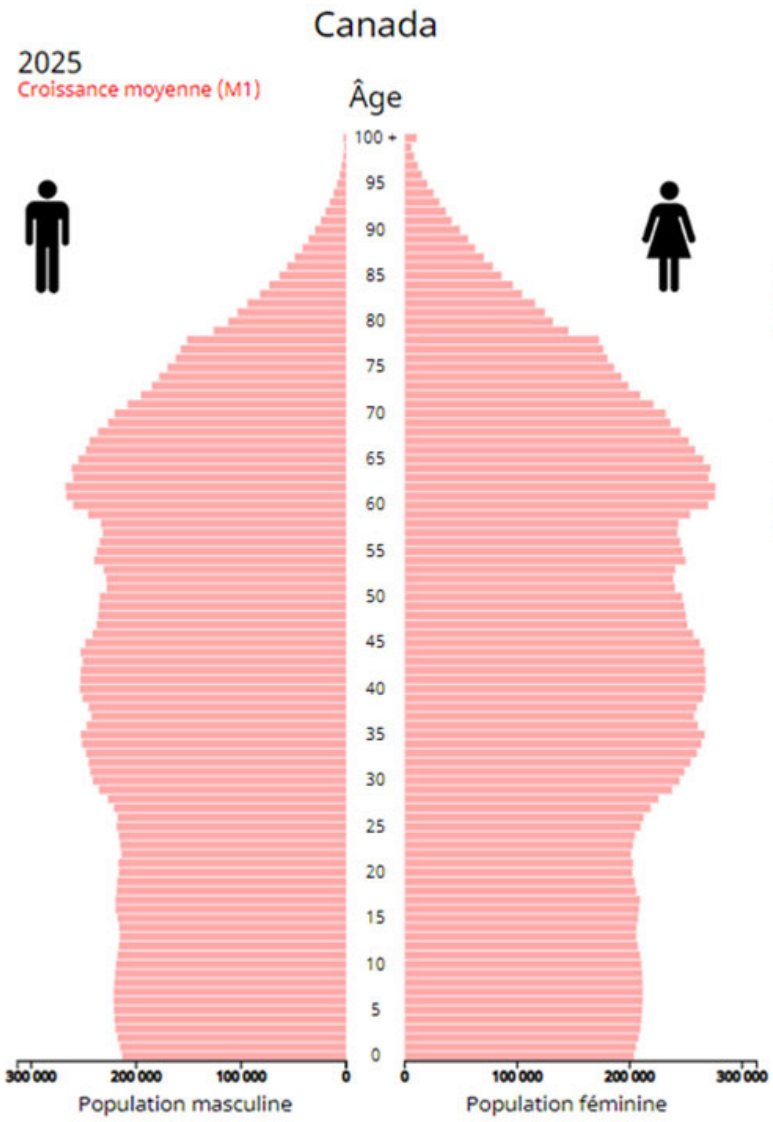
Pyramides des âges

La figure 1.2.1 montre un exemple de pyramide des âges dynamique conçue à partir des projections de la population canadienne pour 2025 selon un scénario de croissance moyenne. La pyramide consiste en deux histogrammes horizontaux placés côte à côte qui indiquent le nombre de personnes par années d'âge. Par convention, l'histogramme des hommes est placé à gauche et celui des femmes à droite.

Voyez le produit Pyramide historique des âges pour d'autres années et scénarios de projection.

Les pyramides des âges sont communément utilisées pour présenter de l'information statistique sur la composition d'une population. Le diagramme montre clairement l'évolution de la génération vieillissante du baby-boom.

Figure 1.2.1
Pyramide des âges de la population canadienne projetée en 2025



Nombre

	0 à 14	15 à 64	65 +
Homme	3 277 775	11 887 465	3 804 240
Femme	3 132 445	12 146 870	4 354 150

Proportion

	0 à 14	15 à 64	65 +
Homme	17,3	62,7	20,1
Femme	16,0	61,9	22,2

Projections (2017 à 2036)

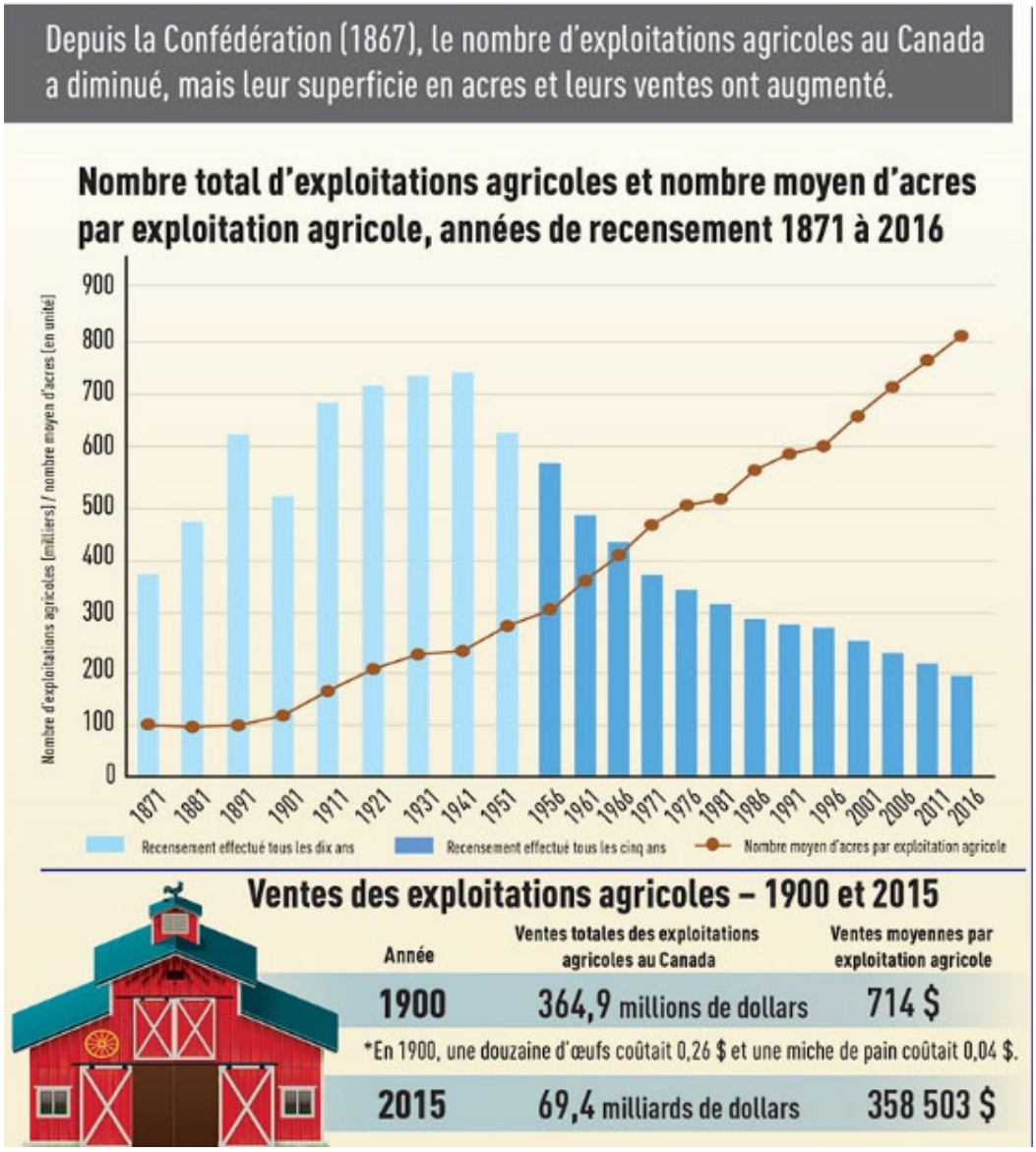
Les différents scénarios de projection sont construits en combinant un certain nombre d'hypothèses concernant l'évolution future de chacune des composantes de la croissance démographique. Ils présentent des séries de résultats distincts mais plausibles en termes d'évolution de la taille et de la structure par âge et sexe de la population afin de refléter l'incertitude associée au futur.

- ○ **Croissance faible**
- ◉ **Croissance moyenne (M1)**
- ○ **Croissance forte**

Infographies

Les infographies combinent des statistiques, des visualisations (comme des graphiques ou des pictogrammes) et du texte pour communiquer une grande quantité d'information statistique de façon compacte et visuellement attrayante. La figure 1.2.2 ci-dessous est un découpage tiré de l'infographie « 150 ans d'agriculture au Canada ». Elle est composée d'un graphique présentant les séries chronologiques du nombre total d'exploitations agricoles et du nombre moyen d'acres par ferme ainsi que d'un tableau indiquant les ventes des exploitations agricoles au Canada en 1900 et 2015.

Figure 1.2.2
Extrait de l'infographie « 150 ans d'agriculture au Canada »



1.3 Qualité des données

De façon générale, l'information statistique est évaluée selon l'adéquation à l'utilisation, c'est-à-dire la mesure dans laquelle l'information statistique répond aux besoins des utilisateurs. À Statistique Canada, l'adéquation à l'utilisation est considérée en fonction de six dimensions.

Dimensions de la qualité et exemples de questions à poser

Pertinence

Est-ce que l'information statistique compte pour les Canadiens?

- Comble-t-elle un besoin d'information?
- Est-elle utile pour bâtir des politiques?
- Aide-t-elle à la planification à long terme?
- Permet-elle de promouvoir de nouvelles initiatives?

Accessibilité

Les utilisateurs ont-ils accès à l'information statistique?

- Est-il facile d'y avoir accès?
- Est-elle bon marché?
- Est-elle organisée et facile à trouver?
- Les utilisateurs peuvent-ils recevoir de l'assistance s'ils rencontrent de la difficulté à y accéder?

Exactitude

L'information statistique est-elle représentative de la mesure qui était ciblée?

- Couvre-t-elle la population ciblée et la période de référence désirée?
- Existe-t-il des sources de sous-dénombrement?
- Les méthodes sont-elles communiquées avec transparence?
- L'information a-t-elle été produite sans influence externe?

Actualité

Le délai entre la période de référence et la disponibilité de l'information statistique est-il acceptable?

- L'information est-elle disponible au moment où l'on en a le plus besoin?
- Accepteriez-vous un moins grand degré d'exactitude pour obtenir l'information plus rapidement?

Intelligibilité

Les métadonnées sont les informations sur les données qui permettent de les mettre en contexte.

- Sont-elles disponibles et complètes?
- Sont-elles utiles?
- Sont-elles fiables?
- Sont-elles disponibles au même moment que les informations auxquelles elles font référence?

Cohérence

Est-ce que l'information statistique est comparable sur une longue période, entre les régions et entre différentes sous-populations?

- Est-ce qu'elle exploite des concepts et des classifications standards?
- A-t-elle été produite avec des méthodes communes à d'autres produits d'information statistique?
- Est-elle comparable à d'autres informations statistiques diffusées antérieurement?

Un compromis est souvent nécessaire entre les dimensions de la qualité. Par exemple, le besoin d'actualité peut avoir un impact sur l'exactitude : diffuser de l'information statistique rapidement réduit le temps disponible pour s'assurer de l'exactitude de l'information.

En plus de la qualité, il est également important de considérer l'éthique des données qui sont collectées et des processus pour produire l'information statistique. Les données de source éthique sont collectées de façon transparente et utilisée d'une manière significative sans causer de tort aux répondants.

1.4 Exercices

1. Identifiez les éléments de la liste qui correspondent à des données structurées et à des données non structurées.

   a. Un relevé bancaire
   b. Un courriel
   c. Une circulaire d'épicerie
   d. Un bulletin scolaire
   e. Les résultats trouvés par un moteur de recherche en ligne

2. Trouvez le terme qui correspond à chacune de ces définitions :

   a. Quelques observations n'ont pas de valeur pour l'une des variables d'un ensemble de données.
   b. Des données qui ont été classées et interprétées.
   c. La valeur la plus représentative des unités d'un ensemble de données.
   d. Un logiciel qui permet de faire des opérations mathématiques sur des données numériques.
   e. Le délai entre la période de référence et le moment où une information est disponible.
   f. Un ensemble de données dans lequel chaque enregistrement correspond à une seule unité d'observation.
   g. Une représentation qui combine des images avec de l'information statistique pour communiquer l'histoire des données.

3. Identifiez la dimension de la qualité qui est en cause dans les situations suivantes :

   a. Vous avez trouvé un ensemble de données parfait pour vos travaux scolaires, mais il y a un coût pour se procurer ces données.
   b. Vous voulez calculer l'âge moyen des personnes dans votre établissement scolaire, mais vous ne connaissez que l'âge des étudiants.
   c. Vous souhaitez explorer un ensemble de données, mais vous ne savez pas à quoi correspondent les variables dans la base de données, car leur nom n'est pas très explicite.
   d. Vous avez fait un sondage dans votre classe pour connaître le niveau d'activité physique des étudiants, mais certains ont répondu par un nombre de pas et d'autres par une distance en kilomètre. Il est donc difficile d'identifier lesquels font le plus d'exercice dans une semaine.

4. Vrai ou faux?

   a. Les statistiques sont les matériaux bruts à partir desquels est produite l'information statistique.

   b. En 2025, il y aura moins de personnes de 15 ans dans la population canadienne que de personnes de 45 ans.

   c. Au deuxième trimestre de 2019, les ventes au détail avaient augmenté dans tous les groupes de marchandise comparativement à l'année précédente.

   d. L'objectif des statistiques est de fournir une information utile au moyen des nombres.

1.5 Réponses

1. a. Structurées
   b. Non structurées
   c. Non structurées
   d. Structurées
   e. Non structurées

2. a. Valeur manquante
   b. Information statistique
   c. Mesure de tendance centrale
   d. Logiciel statistique
   e. Actualité
   f. Microdonnées
   g. Infographie

3. a. Accessibilité
   b. Exactitude
   c. Intelligibilité
   d. Cohérence

4. a. Faux
   b. Vrai
   c. Faux
   d. Vrai

2 Les sources de données

Dans la section 1, nous avons appris que les données sont les matériaux bruts qui sont utilisés pour produire des statistiques et de l'information statistique. Mais d'où viennent exactement ces matériaux bruts et qui est intéressé à les utiliser pour produire de l'information statistique ou à utiliser cette information statistique pour prendre des décisions? Dans cette section, nous discuterons des fournisseurs et des utilisateurs des données, ce qui nous amènera à explorer les différents types de données qui peuvent être utilisés pour produire de l'information statistique ainsi qu'aux avantages et aux inconvénients de chacun de ces types.

2.1 Fournisseurs et utilisateurs des données

Les fournisseurs de données sont des individus ou des organisations qui collectent et traitent les données à des fins diverses. Ils rendent accessibles ces données aux utilisateurs des données qui, à leur tour, les utilisent pour produire de l'information statistique. Les utilisateurs des données peuvent faire partie de la même organisation que les fournisseurs de données ou être une tierce partie. Tout dépendant des raisons pour lesquelles les données ont été collectées à la base, la source de ces données sera dite primaire ou secondaire.

Les données d'une source primaire ont été collectées dans le but de produire des statistiques et de l'information statistique. Les chercheurs, les entreprises et les agences gouvernementales sont les principales sources primaires de données. Les chercheurs et les groupes d'intérêt peuvent obtenir des fonds pour étudier des enjeux sociaux, économiques ou scientifiques, ce qui peut nécessiter la collecte de nouvelle information s'il y a une lacune dans les données, c'est-à-dire que les données sur le sujet ne sont pas disponibles ailleurs. Les entreprises peuvent collecter des données à des fins statistiques dans le cadre de leur vocation principale (firmes de sondage, recherche publicitaire) ou pour des besoins internes (par exemple, sondage auprès des employés). Les agences gouvernementales, par exemple les offices nationaux de statistiques tels que Statistique Canada, ont comme mandat de collecter des données sur différents aspects de la population et de la société pour aider les gouvernements à prendre des décisions et à bâtir des politiques publiques.

Les données d'une source secondaire ont été collectées dans un but autre que celui de produire de l'information statistique. Tout individu ou organisation peut collecter une grande quantité d'information pour plusieurs raisons différentes, comme rémunérer les employés, maintenir des inventaires ou améliorer le fonctionnement d'applications pour téléphone intelligent, par exemple. Ce ne sont pas toutes les sources de données secondaires qui sont adéquates pour produire de l'information statistique, mais les utilisateurs des données sont de plus en plus intéressés à utiliser certaines sources de données secondaires pour en produire. Lorsqu'ils choisissent d'utiliser des données secondaires, les utilisateurs doivent considérer le fournisseur de données et ses motivations pour avoir collecté les données. Ceci les aide à s'assurer que les données sont adéquates pour l'utilisation qu'ils veulent en faire.

Voici quelques groupes et organisations qui utilisent les données, l'information statistique et les statistiques :

- **Administrations publiques :** Les administrations fédérales, provinciales et locales ont besoin de se renseigner notamment sur la population et l'économie, ce qui les aide à élaborer, mettre en œuvre et surveiller les programmes socioéconomiques et environnementaux et d'autres fonctions gouvernementales comme l'attribution de permis et la réglementation. Ces informations permettent aux gouvernements de prendre des décisions sur des enjeux tels que l'emplacement de nouveaux hôpitaux, l'implantation des services ou l'établissement des recettes à tirer de l'assiette fiscale. De même, cela permet au public, aux partis d'opposition et aux groupes d'intérêts de mesurer la performance du gouvernement dans la prise de décisions et dans la justification de sa gestion s'il ne répond pas aux critères.
- **Entreprises :** Les entreprises canadiennes ont besoin de renseignements sur l'économie locale, provinciale et nationale, sur les besoins non comblés des consommateurs et sur les diverses tendances sociales. Elles pourront ainsi mieux décider des mesures d'embauche à prendre, de la mise en marché de leurs produits et de l'implantation de bureaux, d'entrepôts ou d'usines. Les renseignements leur sont aussi nécessaires pour conduire différentes opérations telles que la facturation, la tenue d'inventaire et l'approvisionnement.
- **Groupes communautaires :** Ces organismes ont besoin de renseignements sur une grande variété de sujets comme la santé des peuples autochtones, la répartition de la population, ainsi que le nombre et le

lieu de résidence des immigrants canadiens qui doivent connaître le français ou l'anglais. Les clubs sportifs pourraient vouloir se renseigner sur les assistances aux parties ou le nombre de jeunes dans leur localité.

- **Milieux académiques et chercheurs :** Les données jouent plusieurs rôles pour ceux qui mènent des études et des analyses. Elles peuvent être utilisées pour planifier les recherches (par exemple, décider dans quelle communauté il faudrait effectuer une étude) ou pour soutenir des affirmations et des hypothèses de recherche (par exemple, est-ce que les données historiques soutiennent une corrélation entre le réchauffement climatique et l'augmentation des inondations).
- **Particuliers :** Tous, des élèves aux retraités, ont besoin de certains renseignements à un moment quelconque de leur vie, que ce soit pour terminer une composition, réaliser un grand projet ou encore par simple curiosité.

Les exemples ci-dessus illustrent que les besoins des utilisateurs d'information couvrent l'ensemble du spectre allant des données non traitées jusqu'aux produits d'information statistique finis. C'est-à-dire qu'à des degrés divers, les utilisateurs des données peuvent aussi en être les fournisseurs.

2.2 Types de données

Il existe plusieurs façons de collecter les données, mais les agences comme Statistique Canada ont principalement recours à trois grands types de méthodes: les recensements, les enquêtes-échantillon et les données administratives. Chacun présente à la fois des avantages et des inconvénients qui seront présentés dans cette section. Ensuite, d'autres méthodes alternatives seront décrites.

Recensement

En général, un recensement fait référence à une collecte de données auprès de chaque unité d'un groupe ou d'une population. Si vous aviez recueilli des données sur la taille de tous les élèves de votre classe, ce serait un recensement de votre classe. Les recensements sont souvent utilisés non seulement pour collecter des données à propos des unités d'une population, mais également pour les lister et les dénombrer. Si vous vouliez savoir combien de personnes habitent dans votre rue, vous auriez besoin de faire une liste de tous les logements dans votre rue et ensuite la liste de toutes les personnes qui habitent dans chacun des logements. Ce faisant, vous pourriez décider de collecter d'autres informations comme l'âge, le sexe et la langue maternelle. Ceci vous permettrait de compter le nombre d'hommes, de femmes et d'enfants qui habitent votre rue. Donc un recensement serait une manière directe de dénombrer le nombre d'unités et de produire des statistiques sur différentes caractéristiques.

Voici quelques avantages et désavantages d'utiliser un recensement :

Avantages (+)

Pas de variabilité échantillonnale : Il n'y a pas de variabilité échantillonnale attribuée aux statistiques issues d'un recensement parce qu'elles sont calculées à partir de données sur la population entière.

Fin niveau de détail : Avec un recensement, vous seriez capable de produire des statistiques pour des petits sous-groupes de la population, pourvu que vous ayez collecté les bonnes variables de classification.

Estimation directe des comptes : Le recensement permet une estimation directe des comptes de population, bien que des ajustements puissent être considérés pour les unités qui n'ont pas pu être rejointes.

Inconvénients (-)

Coût élevé : La tenue d'un recensement peut être dispendieuse si la population visée est grande.

Actualité : Un recensement prend plus de temps à réaliser qu'une enquête-échantillon ce qui signifie un plus grand délai entre la date référence et la diffusion des résultats.

Fardeau de réponse élevé : Il faut avoir de l'information sur chacun des membres de la population visée.

Moins de contrôle sur la qualité : Si la taille de la population est beaucoup plus grande que celle d'une enquête-échantillon et que les ressources sont limitées, il se peut que des compromis soient nécessaires sur le plan du contrôle de la qualité. Par exemple, peut-être qu'une partie seulement des non-répondants pourront être rejoints dans le cadre du suivi des cas pour la non-réponse.

Information moins détaillée : Étant donné les coûts, le fardeau de réponse et l'ampleur des activités nécessaires pour conduire un recensement dans une grande population, les variables mesurées sont parfois limitées à une courte liste de variables d'identification et de classification.

Enquête-échantillon

Une enquête peut être n'importe quelle activité de collecte d'information organisée et méthodique à propos des caractéristiques des unités d'une population. À Statistique Canada, les enquêtes utilisent des concepts bien définis ainsi que des méthodes éprouvées qui seront décrites dans la troisième section de ce document. Un recensement peut être considéré comme un type d'enquête, mais le mot **enquête** est le plus souvent utilisé pour faire référence à une enquête-échantillon, c'est-à-dire une enquête où les données sont collectées seulement pour certaines unités d'une population visée. Si vous obtenez la taille de 10 élèves d'une classe de 30 élèves, vous aurez utilisé une enquête-échantillon de votre classe plutôt qu'un recensement.

Voici les avantages et désavantages d'utiliser une enquête-échantillon au lieu d'un recensement:

Avantages (+)

Coût plus bas : Une enquête-échantillon est moins coûteuse qu'un recensement puisque les données sont recueillies auprès d'une partie seulement d'un groupe de la population.

Résultats plus rapides : On obtient des résultats bien plus rapidement que dans un recensement, car il y a moins d'unités à rejoindre et il y a moins de données à traiter.

Fardeau de réponse moins élevé : Moins de gens doivent répondre au questionnaire d'une enquête-échantillon.

Plus de contrôle sur la qualité : La plus petite envergure des activités facilite la gestion et le contrôle de la qualité.

Inconvénients (-)

Variabilité échantillonnale : Si vous sélectionnez plusieurs échantillons d'une même population et calculez des statistiques sur chacun de ces échantillons, les résultats seront un peu différents d'un échantillon à l'autre. Il faut tenir compte de cette source d'incertitude lors de l'estimation des statistiques tirées d'une enquête-échantillon.

Statistiques à un niveau moins détaillé : Il pourrait être impossible de produire des statistiques pour des petites sous-populations ou régions géographiques si elles ne sont pas suffisamment représentées dans l'échantillon.

Données administratives

Les données administratives sont collectées par des organismes dans le cadre de leurs opérations quotidiennes. Ces données portent, par exemple, sur les naissances, les décès, les impôts, les immatriculations de véhicules automobiles ou les transactions. Ces données administratives peuvent être utilisées plus tard à titre de substitut ou en soutien à une enquête-échantillon ou un recensement.

Voici les avantages et désavantages d'utiliser des données administratives plutôt qu'un recensement ou une enquête-échantillon :

Avantages (+)

Coût plus bas : Les données administratives sont moins dispendieuses à utiliser, car il n'y a pas d'opération de collecte.

Pas de variabilité échantillonnale : Il n'y a pas de variabilité échantillonnale attribuée aux statistiques parce qu'elles sont calculées à partir de données sur des groupes entiers de la population.

Séries chronologiques : La collecte de données est continue, d'où la possibilité d'analyser les tendances.

Pas de fardeau de réponse : Il n'y a pas de fardeau additionnel pour les répondants puisque les données sont déjà recueillies.

Fin niveau de détail : Avec les données administratives, vous seriez capable de produire des statistiques pour de petits sous-groupes de la population ou des petites unités géographiques, tant que vous disposez des bonnes variables de classification et que les sous-groupes ont une bonne couverture (c'est-à-dire que la plupart des unités appartenant à ces sous-groupes sont présentes dans le fichier).

Inconvénients (-)

Manque de souplesse : À la différence des données d'enquête, l'utilisateur des données a peu de contrôle sur le choix des variables qui sont collectées. Celles-ci peuvent dans certains cas se limiter à quelques renseignements administratifs essentiels.

Manque d'exhaustivité : Les données se limitent à la population figurant dans les dossiers administratifs. Cette population est souvent différente de la population cible. Plusieurs sources de surdénombrement et de sous-dénombrement sont possibles.

Comparabilité au fil du temps : Les définitions sont conçues à des fins précises et elles évoluent au fil du temps. Ceci peut nuire à la comparabilité si on souhaite étudier des tendances.

Concepts et définitions : Les définitions sont établies par ceux qui conçoivent et gèrent le dossier selon leurs besoins et ces définitions peuvent ne pas être pertinentes dans un autre contexte.

Qualité des données : La qualité des données peut varier d'un fournisseur de données à l'autre, car ils n'accordent pas tous la même importance aux différentes dimensions de la qualité.

Éthique : Avec les recensements et les enquêtes-échantillon, les répondants sont conscients des données qui sont collectées. Ils consentent à ce que ces données soient utilisées puisque la vaste majorité des enquêtes sont faites sur une base volontaire. Avec les données administratives, il serait difficile d'informer chaque personne et d'obtenir son consentement. Ceci implique que les individus et les organisations qui utilisent les données administratives pour produire de l'information statistique ont une grande responsabilité de s'assurer que les données sont utilisées d'une manière bénéfique pour la société et que l'éthique des données a été considérée à toutes les étapes du processus.

Sources de données alternatives

Ces sources de données sont de plus en plus utilisées dans la production d'information statistique pour remplacer ou compléter les méthodes traditionnelles.

L'**approche participative** consiste à recueillir des renseignements provenant d'une vaste communauté d'utilisateurs et repose sur le principe selon lequel chaque citoyen est un expert dans son milieu. Avant la légalisation du cannabis en 2018, le gouvernement canadien avait besoin d'information sur la taille et l'activité du marché noir existant pour le cannabis séché. Cette information était difficile à collecter par une enquête-échantillon traditionnelle. D'une part, la caractéristique mesurée était rare. Un échantillon probabiliste aurait inclus plusieurs personnes qui ne consomment pas de cannabis puisque ceux-ci sont plus nombreux que les

consommateurs dans la population canadienne. D'autre part, certaines personnes auraient pu hésiter à donner les détails de leur consommation de cannabis à un intervieweur. Statistique Canada a alors opté pour une approche participative pour recueillir l'information. L'agence a établi StatsCannabis, une application web anonyme permettant aux consommateurs de rapporter l'information sur leurs achats. Le gouvernement a pu utiliser cette information pour planifier la transition vers un marché légal du cannabis.

Le **moissonnage du web** est un processus par lequel des renseignements sont recueillis et copiés à partir du web aux fins d'analyses ultérieures. Depuis janvier 2021, des données du moissonnage du web sont utilisées pour modéliser le prix des ordinateurs dans l'Indice des prix des ordinateurs, des logiciels et des fournitures informatiques, une composante de l'Indice des prix à la consommation. Ce changement de la méthode de collecte vise à améliorer la couverture des produits considérés et l'actualité de l'information sur leurs prix, considérant les changements rapides propres à l'économie numérique. Comme pour les données administratives, les utilisateurs des données moissonnées ont une responsabilité accrue de s'assurer de l'éthique des données collectées et de suivre les meilleures pratiques pour éviter de collecter des informations personnelles par inadvertance.

La **télédétection** est l'acquisition à distance de renseignements à propos d'un objet ou d'un phénomène. La télédétection est utilisée à Statistique Canada pour le Programme d'évaluation de l'état des cultures. La croissance végétale sur les fermes canadiennes est observée de façon hebdomadaire à l'aide de l'imagerie satellite. Les données sont généralement traitées et rendues disponibles le même jour, permettant un monitorage en temps réel de l'agriculture canadienne. Ce programme fournit de l'information de grande valeur tout en réduisant les coûts de collecte et le fardeau de réponse des producteurs agricoles. D'autres exemples de télédétection sont les radars météorologiques qui suivent les tempêtes et les sismographes qui mesurent les vibrations de la terre.

Les **registres statistiques** sont des ensembles de données créés à des fins statistiques qui sont continuellement mises à jour avec des renseignements sur toutes les unités d'une population. Ils sont souvent créés par l'intégration de multiples sources de données à l'aide du couplage des microdonnées et utilisent des algorithmes ou des techniques d'apprentissage automatique pour consolider l'information et dériver de nouvelles variables. Le Registre des entreprises de Statistique Canada est un exemple de registre statistique qui est mis à jour en continu à partir des données sur les taxes payées par les entreprises et des données d'enquête. Il sert de base de sondage pour un grand nombre d'enquêtes économiques et il permet de produire les comptes semi-annuels des entreprises.

Finalement, les **données ouvertes** et les **mégadonnées** sont d'autres termes utilisés pour décrire certains types de données. Les données ouvertes sont des données structurées, directement exploitables par un ordinateur, qui sont partagées gratuitement et qui peuvent être utilisées sans restriction. Les mégadonnées réfèrent à des ensembles de données dont le nombre d'enregistrements et le nombre de variables sont si élevés qu'ils dépassent les capacités des logiciels traditionnels à traiter l'information en un temps raisonnable. Elles sont aussi caractérisées par les trois « v » : volume, variété et vélocité.

2.3 Exercices

1. Quels éléments (par exemple, le coût) devriez-vous considérer au moment de choisir une méthode de collecte de données?

2. Quel type de collecte de données (recensement, enquête-échantillon ou utilisation de données administratives) serait le meilleur moyen de répondre aux questions ci-dessous :

a. Quelle est la cause principale de décès des jeunes Canadiens âgés de 15 à 25 ans?

b. Quel type d'aliments devrait-on commander pour un pique-nique de classe, en tenant compte des préférences des élèves?

c. Le directeur d'une compagnie de téléphonie cellulaire se pose la question suivante : si on offrait de nouveaux services, comment réagiraient nos clients actuels?

2.4 Réponses

1. Lorsque vous choisissez une méthode de collecte de données, vous devez considérer les éléments suivants :

- le coût (budget)
- l'échéance
- la taille de la population
- le personnel nécessaire pour appliquer la méthode choisie

2. Quel type de collecte de données (recensement, enquête-échantillon ou utilisation de données administratives) serait le meilleur moyen de répondre aux questions ci-dessous.

a. Utilisation de données administratives : les renseignements seraient fournis par un registre des décès.

b. Un recensement, parce qu'une classe comporte un nombre raisonnable d'élèves à questionner.

c. Une enquête-échantillon : cela serait plus rentable et prendrait moins de temps qu'un recensement de tous les clients.

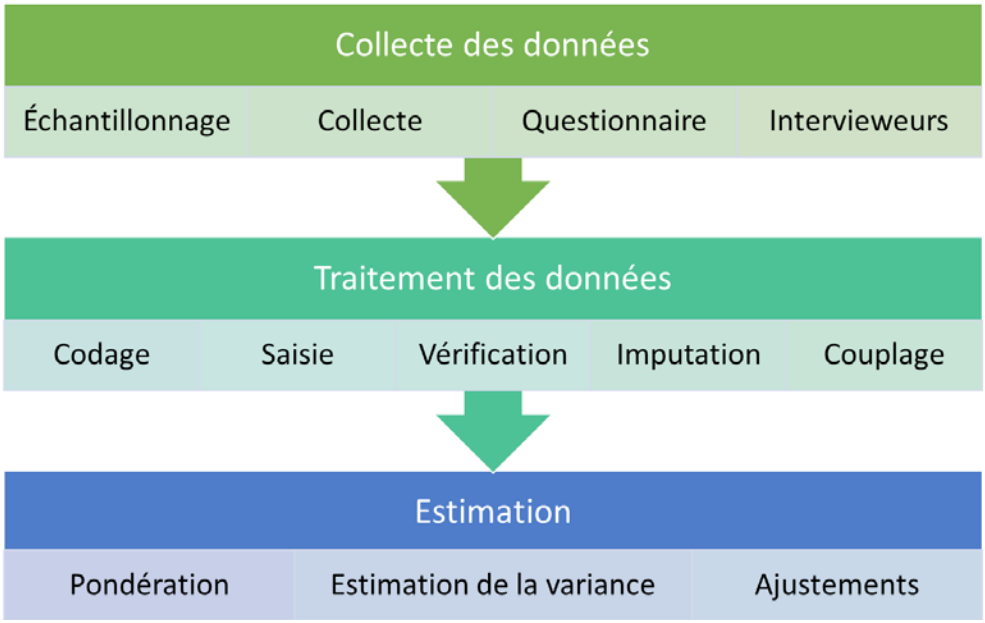
3 Collecte et traitement des données

La section 2 décrivait les acteurs de la société qui collectent et utilisent des données et à quelles fins. Ensuite, les avantages et désavantages des différents types de méthode de collecte traditionnelle ont été comparés. Finalement, des exemples de différents types de données alternatives récemment utilisées à Statistique Canada ont été donnés.

Parmi ces méthodes de collecte, les enquêtes-échantillons sont les plus fréquemment utilisées dans la production d'information statistique à propos de la société canadienne. Elles requièrent toutefois une planification minutieuse et des méthodes éprouvées pour la sélection d'échantillons et la tenue des activités de collecte. Par conséquent, les trois premières sous-sections de la section 3 sont consacrées aux enquêtes-échantillons tandis que les autres sous-sections sur le traitement, l'estimation et la gestion de la qualité peuvent potentiellement concerner tous les types de données.

La figure 3.1 résume les différentes étapes de la collecte des données, du traitement des données et de l'estimation qui seront explorés dans cette section.

Figure 3.1 Les différentes étapes de la collecte des données, du traitement des données et de l'estimation



3.1 Planification

A priori, le déroulement d'une enquête pourrait simplement consister à poser des questions et à compiler les réponses pour obtenir des statistiques. Il faut cependant suivre des étapes précises afin que les résultats de l'enquête fournissent de l'information fiable et utile.

Au départ, les questions suivantes doivent être discutées :

- Pourquoi mène-t-on cette enquête?
- Qui fait l'objet de l'information à recueillir?
- Qu'ai-je besoin de savoir?
- Comment l'information sera-t-elle utilisée?
- Quel degré d'exactitude et de fiabilité de l'information doit-on viser?

Pour réaliser un plan d'enquête, il faut prendre une foule de décisions sur les sujets ci-dessous, qui seront expliqués de façon plus détaillée dans la présente section.

- Objectifs de l'enquête
- Population cible
- Besoins en données
- Choix du type de collecte
- Minimiser l'erreur
- Taille de l'échantillon
- Plan d'analyse
- Conception du questionnaire
- Méthodes de collecte de données
- Plan de traitement de données
- Contrôle de la qualité
- Analyse et diffusion des résultats

Objectifs de l'enquête

Le plan d'enquête commence par des objectifs qui décrivent pourquoi et auprès de quelle population l'enquête doit être réalisée. Ces objectifs en disent long sur les données qui devront être recueillies. Ils aident aussi à cibler la population cible.

Exemple

Imaginons que le conseil étudiant de l'École secondaire Saint-Sauveur décide d'enquêter auprès des élèves afin d'obtenir de l'information pour les aider à planifier le bal des finissants. À partir de ce but général, on peut raffiner les objectifs. Disons que les objectifs sont les suivants :

- Recueillir de l'information auprès des élèves afin de déterminer les facteurs qui feront du bal des finissants un succès. (On définit un succès d'après les critères suivants : le plus grand nombre d'élèves possibles viendront au bal et celui-ci comblera leurs attentes.)
- Obtenir des données utiles qui serviront au comité organisateur du bal des finissants.

Le plan d'enquête démontre comment les objectifs de l'enquête seront atteints en décrivant clairement la population cible, les besoins en données et les variables qui seront mesurées, tout en prévoyant les questions et réponses possibles, ainsi que la manière dont les données seront traitées et analysées.

Population cible

Si l'objectif d'une enquête est d'obtenir de l'information auprès des élèves, par exemple, il faudra se demander « de quels élèves s'agit-il? » afin de définir la population cible.

Dans l'exemple décrit plus haut, le comité organisateur du bal voudra probablement consulter seulement les élèves qui prévoient être diplômés cette année, donc ceux qui fréquentent la dernière année du secondaire (secondaire 5 ou 12e année). Si certains d'entre eux étudient à temps partiel et ne prévoient pas terminer cette année, ils n'auront pas besoin d'être consultés. La population cible serait donc définie comme « les élèves de l'École secondaire Saint-Sauveur qui fréquentent le secondaire 5 à temps plein ».

Il arrive parfois que la population cible, c'est-à-dire la population au sujet de laquelle on souhaite recueillir de l'information, et la population observée, c'est-à-dire la population réellement couverte par l'enquête, soient différentes pour des raisons pratiques. Idéalement, les deux populations devraient être très similaires. Il est important de noter que les conclusions de l'enquête ne pourront être généralisées qu'à la population observée.

Dans notre exemple, il se peut que certains élèves de secondaire 5 à temps plein soient absents de l'école au moment de l'enquête. Puisqu'il serait trop difficile de rejoindre ces élèves pour les consulter, ils ne feraient donc pas partie de la population qui serait interrogée, même s'ils font partie de la population cible.

De plus, il est possible que certains concepts ou méthodes d'enquête ne soient pas appropriés pour certains segments de la population. Prenons par exemple une enquête sur les diplômés d'études postsecondaires dont l'objectif serait de déterminer s'ils ont trouvé un emploi et, le cas échéant, quel genre d'emploi ils ont décroché. Dans un tel cas, on pourrait exclure de la population cible les diplômés qui ont étudié dans des écoles spécialisées telles que les collèges militaires. Ces types de diplômés seraient presque certains d'obtenir un emploi dans leur domaine. Ainsi, la population cible pourrait comprendre seulement ceux qui ont obtenu un diplôme d'une université, d'un collège ou d'une école professionnelle.

Il peut aussi s'avérer nécessaire d'imposer des limites géographiques en omettant certaines parties de la population cible. Certaines régions pourraient être inaccessibles en raison des coûts de déplacement élevés ou autres difficultés. On peut imaginer, par exemple, qu'une entreprise faisant une enquête au moyen d'entrevues en personne voudrait tirer un échantillon dans des régions densément peuplées afin de minimiser les déplacements.

Besoins en données

Afin de définir les besoins en données, il faut se poser les questions suivantes : « Que veut-on savoir au juste? » et « Comment se servira-t-on de l'information recueillie? ».

Dans notre exemple, voici quelques questions que le comité organisateur pourrait considérer :

- Faut-il savoir le nombre de personnes qui ont l'intention d'aller au bal? (On pourrait aussi le savoir d'après le nombre de billets vendus.)
- Si l'on décide de demander aux élèves s'ils ont l'intention d'aller au bal, devrait-on poser des questions particulières à ceux qui ne prévoient pas y aller? En comprenant mieux leurs raisons de se désister, il serait possible de planifier des activités qui les intéressent, ce qui pourrait leur faire changer d'idée!
- Pour connaître les préférences des élèves en ce qui concerne le bal, quels aspects veut-on considérer? Sans doute des éléments comme :
  - le coût du billet,
  - la musique,
  - le type de nourriture et de breuvages,
  - le jour de la semaine,
  - l'emplacement.
- Y a-t-il d'autres facteurs à considérer? Les élèves veulent-ils un photographe? Est-ce que tous préfèrent un repas avant le bal dansant, ou certains veulent-ils uniquement un bal?
- Les élèves voudraient-ils avoir des gardes de sécurité à l'entrée du bal? Quel moyen de transport prévoient-ils utiliser pour s'y rendre? (On pourrait considérer la location d'un autobus qui partirait d'un endroit central.)

Lorsque l'on planifie le contenu d'une enquête, il est tentant de vouloir recueillir le plus d'information possible. Cependant, plus il y a de questions, plus l'enquête prendra du temps et plus elle coûtera cher. Il faut se demander « A-t-on vraiment besoin de cette information? » tout en tenant compte du temps et des ressources nécessaires pour tester le questionnaire, traiter les données et analyser les résultats.

De plus, il faut tenir compte du fardeau de réponse afin de ne pas trop incommoder les répondants. Dans les enquêtes, le fardeau de réponse est affecté par :

- le nombre de questions posées,
- la nature délicate des questions posées,
- le nombre de fois qu'on communique avec le répondant (lors d'une même enquête ou lors de plusieurs enquêtes),
- le détail du renseignement demandé (par exemple, si on demande un revenu précis, le répondant doit consulter ses documents officiels, mais si on demande plutôt de choisir entre cinq tranches de revenu, il peut répondre plus facilement),
- le temps qu'il faut pour remplir le questionnaire.

Choix du type de collecte

Le degré de précision et de détails de l'information que l'on vise à collecter ainsi que les ressources dont on dispose influencent le choix du type de collecte. Les avantages et inconvénients de chaque type ont été discutés précédemment dans la section sur les types de données.

Dans notre exemple, le comité organisateur peut choisir de faire un recensement de tous les finissants ou d'interroger un échantillon seulement de ce groupe.

Le type de collecte est souvent déterminé en fonction du **budget** alloué pour mener l'enquête. Les coûts représentent une des principales raisons de mener une enquête-échantillon au lieu d'un recensement. Grâce à l'échantillonnage, il est possible d'obtenir des résultats valables au moyen d'un échantillon relativement petit de

la population cible. Par exemple, s'il faut obtenir des données sur tous les citoyens canadiens de 15 ans et plus, une enquête menée auprès d'un petit nombre de ces derniers (1 000 ou 2 000 personnes selon les besoins en données) pourrait fournir des résultats satisfaisants.

Un autre avantage de l'enquête-échantillon est qu'il permet de produire des données peu de temps après qu'un besoin d'information ait été identifié, selon un **échéancier** rapide. Par exemple, si un organisme veut mesurer le degré de sensibilisation du public à la suite de sa campagne publicitaire, il doit mener une enquête peu de temps après la réalisation de sa campagne. En utilisant un échantillon de la population cible, la durée de la collecte et du traitement des données est diminuée et davantage de temps est consacré à la planification de l'enquête et au contrôle de qualité.

Minimiser l'erreur

Lors de la planification, il faut prévoir les sources potentielles d'erreurs afin de les réduire le plus possible.

Dans une enquête-échantillon, la variation qui existe parmi différents échantillons cause une certaine incertitude qu'on appelle l'erreur due à l'échantillonnage. Par exemple, disons que vous estimez la distance moyenne entre la maison et l'école des élèves d'une classe de 25 élèves, à partir d'un échantillon de 5 élèves. Votre estimation dépendra de l'identité des 5 élèves échantillonnés. Si ceux-ci vivent tous très près de l'école, les résultats ne seront pas représentatifs de l'ensemble de la classe. C'est la variation d'un échantillon à l'autre qui cause l'erreur d'échantillonnage.

En général, plus il y a de personnes dans l'échantillon, plus l'erreur due à l'échantillonnage sera petite. Il est souvent possible d'estimer l'erreur associée à différentes méthodes d'échantillonnage pour ensuite tenter de la réduire.

En choisissant de faire un recensement, l'erreur liée à la variation dans l'échantillon est évitée, mais pas les autres sources d'erreurs, appelées les erreurs non dues à l'échantillonnage. Par exemple, une question pourrait être posée de façon à encourager une certaine réponse ou bien des erreurs pourraient survenir en traitant les données ou en calculant un pourcentage pour un tableau de données. Il faut éviter le plus possible ce type d'erreurs en portant attention au contrôle de qualité durant toutes les étapes de l'enquête.

Les deux types d'erreurs seront discutés plus en détail dans la sous-section consacrée à l'estimation.

Taille de l'échantillon

Puisque les enquêtes-échantillons sont toutes différentes, il n'existe pas de règles rigoureuses et universelles pour déterminer la taille d'un échantillon. Les facteurs décisifs sont l'échéancier, les coûts, les contraintes opérationnelles et la précision désirée des résultats. En évaluant chacun de ces facteurs, il est plus facile de prendre une décision par rapport à la taille de l'échantillon. De plus, il faut tenir compte du niveau acceptable d'erreur d'échantillonnage. S'il existe une caractéristique en particulier à mesurer qui est primordiale dans l'enquête et que des variations considérables de cette caractéristique sont observées au sein de la population, la taille de l'échantillon devra être plus grande afin d'obtenir une meilleure précision.

Plan d'analyse

Après avoir identifié tous les éléments ou variables à mesurer et préparé le plan d'échantillonnage, la prochaine étape consiste à élaborer le plan d'analyse. Il s'agit de concevoir les tableaux qu'il sera possible de produire à partir des variables de l'enquête. Ces tableaux ne contiendront pas encore de données, mais ils montreront les croisements proposés entre les variables.

> Dans notre exemple, le comité organisateur pourrait concevoir des tableaux de résultats pour chaque variable prévue, présentés sous forme de nombre et de pourcentage (par exemple le nombre et le pourcentage d'élèves ayant préféré l'emplacement A et l'emplacement B pour le bal). Certains tableaux pourraient aussi présenter des croisements comme « La musique préférée selon le genre ».

Ces tableaux aident à vérifier si les questions prévues permettent de répondre aux objectifs de l'enquête. Ils illustrent de façon concrète comment l'information recueillie pourra être utilisée et comment elle constitue une réponse adéquate aux besoins en données.

Conception du questionnaire

Le questionnaire est développé à partir des besoins en données précédemment identifiés et du plan d'analyse qui a été développé. Pour formuler les questions, il est utile de consulter les utilisateurs des données et les experts en la matière ou d'examiner les questions d'autres enquêtes portant sur le même sujet ou sur des thèmes semblables. Les bonnes pratiques en matière de conception et de mise à l'essai des questionnaires seront discutées dans la sous-section consacrée à la collecte des données.

Méthodes de collecte de données

Le choix de la méthode pour recueillir les données aura des conséquences directes sur les coûts, les ressources matérielles et humaines, le temps requis pour mener l'enquête et pour évaluer la qualité des données. Une première option est l'entrevue, qui peut être faite sur place ou au téléphone, avec ou sans ordinateur. Une seconde option est le questionnaire à remplir soi-même, en format papier ou en format électronique.

L'entrevue en personne, qui est menée par un intervieweur qualifié, peut comporter des techniques de questionnement structuré ou non structuré. Lorsque l'entrevue se déroule au téléphone, les questions sont structurées au moyen d'un protocole formel.

Le questionnaire à remplir soi-même doit être très structuré puisque le répondant ne pourra pas obtenir autant d'aide qu'avec d'un intervieweur. Il retournera le questionnaire par la poste, par un autre mode ou encore il le complétera en ligne.

> Dans notre exemple, le comité organisateur pourrait opter pour l'entrevue en personne assistée par ordinateur. Elle serait menée par des intervieweurs qui utiliseraient un ordinateur portable pour saisir les réponses des élèves à l'aide d'un questionnaire électronique. Si certains élèves s'inquiétaient en ce qui concerne la confidentialité de leurs réponses, l'intervieweur pourrait leur donner l'option de saisir eux-mêmes leurs réponses. Cependant, cette option pourrait causer un plus grand nombre d'erreurs et compromettre la qualité des données recueillies, ce qui augmenterait le temps consacré au traitement de données.

Plan de traitement de données

Il s'agit du processus servant à convertir les réponses du questionnaire en données de sortie. Les tâches à accomplir durant l'étape de traitement des données sont le codage, la saisie et la mise en forme des données, la vérification, l'imputation et la construction de variables dérivées. Bref, l'objectif de cette étape est de produire un fichier de données sans valeur invalide ou manquante qui pourra être utilisé pour l'estimation et l'analyse des données.

Contrôle de la qualité

Ce processus sert à repérer les erreurs et à vérifier les résultats. Malgré une planification et des essais rigoureux, il est impossible de prévoir toutes les difficultés d'une enquête. C'est pourquoi aucune enquête n'est parfaite. Les tâches de contrôle de la qualité sont nécessaires afin de réduire les erreurs non dues à l'échantillonnage qui se sont glissées à chaque étape de l'enquête. Ces tâches de contrôle incluent la formation des intervieweurs, l'essai des systèmes informatiques, le suivi auprès des non-répondants et la vérification des données recueillies et des données de sortie. Les programmes de contrôle statistique de la qualité visent à ce que le taux d'erreur soit réduit au minimum.

Analyse et diffusion des résultats

Après la collecte et le traitement des données, il faut prévoir les étapes de l'analyse et de la diffusion des résultats de l'enquête, c'est-à-dire :

- organiser les données à l'aide de tableaux de distribution de fréquences,
- résumer les données au moyen des mesures de tendance centrale et de dispersion,
- présenter les données au moyen de différents types de diagrammes,
- rédiger les conclusions de l'enquête pour ensuite les présenter au public et les publier.

Dans notre exemple, les membres du comité organisateur du bal pourraient se partager les tâches d'organisation et d'analyse des données et de rédaction des conclusions de l'enquête. Les décisions concernant le lieu du bal, le prix du billet, le type de musique, etc. seraient alors fondées sur ces conclusions. En publiant les faits saillants de l'enquête dans le journal de l'école, le conseil étudiant pourrait montrer que ses décisions sont basées sur les attentes exprimées par l'ensemble des élèves.

Les sections sur l'exploration et la visualisation des données présenteront certaines de ces étapes plus en détail.

3.2 Échantillonnage

Plusieurs étapes de la planification d'une enquête reposent avant tout sur une bonne compréhension des besoins des utilisateurs de l'information et une bonne connaissance des étapes à suivre pour réaliser l'enquête. L'une des premières étapes qui nécessitent en plus des connaissances de la statistique mathématique est l'échantillonnage, c'est-à-dire la méthode utilisée pour sélectionner un échantillon de la population visée. Il y a plusieurs méthodes d'échantillonnage possibles et la méthode choisie aura un impact direct sur l'exactitude des statistiques produites. Il est donc important de bien comprendre les différences entre les différents plans d'échantillonnage possibles. Ceux-ci se séparent en deux types : l'échantillonnage probabiliste et l'échantillonnage non probabiliste. Mais d'abord, voyons quels éléments doivent être pris en compte pour choisir le plan le plus adéquat dans un contexte donné.

3.2.1 Sélection d'un échantillon

L'échantillonnage permet d'estimer des caractéristiques d'une population en observant directement une partie de la population. Les chercheurs ne s'intéressent pas à l'échantillon lui-même, mais à ce qu'il leur permet d'apprendre sur l'ensemble de la population. L'enquête-échantillon doit être correctement définie et organisée. Si on pose les mauvaises questions, l'information recueillie ne permettra pas de répondre aux objectifs de l'enquête. Si on pose les questions aux mauvaises personnes, l'information ne représentera pas bien la population à laquelle on s'intéresse. Les résultats seront biaisés.

Voici les étapes à suivre pour sélectionner un échantillon et vous assurer qu'il vous permettra de répondre aux objectifs de l'enquête.

Établir les objectifs de l'enquête

Clarifier les objectifs de l'enquête de façon aussi détaillée que possible est essentiel à son succès définitif. Il faudrait à ce stade identifier les utilisateurs initiaux et définir les utilisations initiales des données. C'est aussi à cette étape que l'on devrait déterminer le type de données le plus approprié à employer parmi le recensement, l'enquête-échantillon, les données administratives ou une source de données alternatives.

Définir la population cible

Peu importe le type de données choisi, il faut bien définir la population cible, c'est-à-dire la population totale pour laquelle on a besoin d'information. Pour ce faire, il faut décrire les unités qui composent la population sous forme de caractéristiques les identifiant clairement. Les caractéristiques suivantes définissent la population cible :

- La nature des unités : des personnes, des hôpitaux, des écoles, etc.
- L'emplacement géographique : il faut déterminer les limites géographiques qui circonscrivent la population et le degré de détail géographique dont on a besoin pour l'estimation découlant de l'enquête (par province, par ville, etc.).
- La période de référence : la période visée par l'enquête.
- D'autres caractéristiques, comme des caractéristiques sociodémographiques (un groupe d'âge particulier, par exemple) ou le type d'industrie.

Déterminer les données à recueillir

Il faut établir les besoins en données. Il faut aussi définir les termes relatifs aux données et s'assurer que ces définitions répondent aux besoins sur le plan opérationnel.

Fixer le degré de précision

Il y a un degré d'incertitude associé aux estimations établies à partir d'un échantillon. Il s'agit de l'erreur d'échantillonnage. Lors de la conception de l'enquête, il faut établir le degré acceptable d'incertitude des estimations à produire. Ce degré dépend de l'utilisation finale des résultats et du budget global de l'enquête. Plus le budget de l'enquête sera élevé, plus on disposera de ressources pour contrôler la qualité. La taille de l'échantillon déterminera aussi le degré d'incertitude. L'accroissement de la taille de l'échantillon entraîne une diminution de l'erreur d'échantillonnage. Par exemple, si vous échantillonnez 24 des 25 élèves de votre classe, il n'y aura pas autant de variation d'un échantillon à un autre qu'il y en aurait si vous n'échantillonniez que 5 élèves parmi les 25 élèves de la classe.

Le plan d'échantillonnage

Les étapes suivantes permettent de définir le plan d'échantillonnage :

1. Déterminer ce que sera la population observée (par exemple, des élèves, des hommes de 20 à 35 ans, des nouveau-nés, etc.).
2. Choisir le délai d'exécution de l'enquête le plus approprié.
3. Définir les unités d'enquête.
4. Établir la taille de l'échantillon (par exemple, un échantillon de 100 pour une population de 1 000).
5. Sélectionner une méthode d'échantillonnage.

Les techniques d'estimation à utiliser, c'est-à-dire la façon dont les résultats seront généralisés à l'ensemble de la population et dont l'erreur d'échantillonnage sera calculée, découleront directement du plan d'échantillonnage et seront décrites dans la section à venir sur l'estimation.

La population observée

Certaines unités de la population cible doivent être exclues en raison de contraintes opérationnelles comme le coût élevé de la collecte des données dans certaines régions éloignées, la difficulté d'identifier et de contacter les personnes appartenant à certains groupes, etc. La population qui est réellement prise en compte pour l'enquête est alors la **population observée**. Bref, la population cible est la population que nous **voulons observer**, tandis que la population observée est la population que nous **pouvons observer**.

Il faut faire en sorte que la population observée se rapproche autant que possible de la population cible. Il est également très important d'informer les utilisateurs des données des différences entre les deux populations, étant donné que les résultats de l'enquête ne s'appliqueront qu'à la population observée.

Par exemple, la population cible d'une enquête pourrait se composer de tous les Canadiens de 15 ans et plus (à une date de référence particulière) tandis que la population observée pourrait exclure les résidents du Yukon, du Nunavut et des Territoires du Nord-Ouest, les personnes vivant sur des réserves autochtones, les membres à temps plein des Forces armées canadiennes et les personnes vivant dans un établissement hospitalier ou carcéral. Ces Canadiens pourraient être exclus pour diverses raisons : parce que sonder des gens dans les territoires pourrait s'avérer difficile et coûteux, parce que le personnel militaire risque de ne pas être disponible à des fins d'enquête s'il est en mission, etc. Dans cet exemple, environ 2 % de la population cible serait exclue de la population observée.

La base de sondage

La base de sondage est l'outil utilisé pour avoir accès à la population. Il existe deux types de bases de sondage : les bases liste et les bases aréolaires. Une base liste est simplement une liste des unités de la population. Chaque unité y est identifiable et la base contient des informations permettant d'accéder aux unités. Une bonne base de sondage devrait être complète et à jour. Aucun membre de la population observée ne devrait en être exclu ni y apparaître plus d'une fois et aucune unité ne faisant pas partie de la population (comme une personne décédée) ne devrait y être inscrite. Le choix de la base de sondage aura des répercussions directes sur la définition de la population observée. Par exemple, si une liste de numéros de téléphone est utilisée pour sélectionner un échantillon de ménages, tous les ménages n'ayant pas le téléphone seront alors exclus de la population observée.

La base aréolaire est une liste d'aires géographiques. Plutôt que de sélectionner directement les unités comme avec une base liste, certaines aires géographiques de la base aréolaire sont sélectionnées et un moyen d'accéder aux unités situées dans ces aires géographiques est identifié, comme visiter ces unités en personne par exemple. Supposons que vous êtes en train d'étudier une ville du Québec située en milieu rural pour déterminer quel pourcentage de ses résidents sont des exploitants agricoles. L'échantillon de la base aréolaire pourrait vous permettre de localiser les routes où rendre visite à des gens, mais vous devriez quand même trouver les noms et les adresses des personnes domiciliées sur chacune de ces routes.

Les unités d'enquête

Il existe trois types d'unités qu'il faut identifier correctement afin d'éviter des problèmes durant la sélection, la collecte et l'analyse des données. Ces unités sont :

- L'unité d'échantillonnage, qui fait partie de la base de sondage et donc qui peut être sélectionnée.
- L'unité déclarante, qui fournit l'information demandée dans le cadre de l'enquête.
- L'unité de référence, ou l'unité d'analyse, qui est l'unité au sujet de laquelle de l'information est fournie et qui sert à analyser les résultats de l'enquête.

Par exemple, dans le cadre d'une enquête sur les nouveau-nés à Edmonton, l'unité d'échantillonnage pourrait être un ménage, l'unité déclarante, l'un des parents ou le tuteur légal, et l'unité de référence, le bébé.

Les unités d'échantillonnage peuvent différer selon la base de sondage utilisée. C'est pourquoi la population observée, la base de sondage et les unités d'enquête sont définies les unes par rapport aux autres.

La taille de l'échantillon

Le degré de précision nécessaire pour les estimations découlant de l'enquête aura des répercussions sur la taille de l'échantillon. Il n'est toutefois pas aussi facile de déterminer la taille de l'échantillon qu'on pourrait le penser. En règle générale, la taille réelle de l'échantillon d'une enquête est un compromis entre le degré de précision à atteindre, le budget de l'enquête et toutes les autres contraintes opérationnelles. Pour atteindre un certain degré de précision, la détermination de la taille de l'échantillon doit reposer entre autres choses sur les facteurs suivants :

- La variabilité des caractéristiques à observer. Si toutes les personnes d'une population gagnaient le même salaire, un échantillon d'une seule personne serait alors tout ce dont vous auriez besoin pour estimer le salaire moyen de la population en question. Si les salaires sont très différents d'une personne à l'autre, vous auriez alors besoin d'un échantillon plus grand pour en produire une estimation fiable.
- La taille de la population : Dans une certaine mesure, plus la population est grande, plus l'échantillon doit être grand. Une fois une certaine taille de population atteinte, une augmentation de la population n'a plus d'influence sur la taille de l'échantillon. La taille de l'échantillon nécessaire pour atteindre un certain degré de précision, par exemple, sera à peu près la même pour une population d'un million que pour une population deux fois plus importante.
- Les méthodes d'échantillonnage et d'estimation : Toutes les méthodes d'échantillonnage et d'estimation ne sont pas aussi efficaces les unes que les autres. Plus la méthode est efficace, moins l'échantillon requis pour obtenir une certaine précision est grand. En raison des contraintes opérationnelles et des limites de la base de sondage utilisée, il se peut que la technique la plus efficace ne puisse pas être utilisée.

La méthode d'échantillonnage

Il existe deux types de méthodes d'échantillonnage : l'échantillonnage probabiliste et l'échantillonnage non probabiliste. La différence entre les deux tient au fait que dans le cas de l'échantillonnage probabiliste chaque unité a une probabilité d'être sélectionnée qui peut être quantifiée, ce qui n'est pas vrai pour l'échantillonnage non probabiliste. Les deux sections suivantes décrivent en détail les méthodes de chaque type.

3.2.2 Échantillonnage probabiliste

L'échantillonnage probabiliste fait référence à la sélection d'un échantillon d'une population lorsque cette sélection repose sur le principe de la randomisation, c'est-à-dire la sélection au hasard ou aléatoire. Il est plus complexe, plus long à mettre en œuvre et habituellement plus dispendieux que l'échantillonnage non probabiliste. Toutefois, comme les unités de la population sont sélectionnées au hasard et qu'il est possible de calculer la probabilité de sélection de chaque unité dans l'échantillon, il permet de produire des estimations fiables et de faire des inférences statistiques au sujet de la population.

Il existe plusieurs méthodes d'échantillonnage probabiliste. Le choix d'un type d'échantillonnage repose sur plusieurs facteurs comme la précision des estimations désirée, la nature de la population d'intérêt, l'information

connue sur cette population de même que des contraintes opérationnelles. Certaines contraintes opérationnelles peuvent aussi influencer ce choix, comme les caractéristiques de la base de sondage.

Dans cette section, les méthodes d'échantillonnage probabiliste seront décrites brièvement et illustrées à l'aide d'exemples.

Échantillonnage aléatoire simple

Dans un **échantillonnage aléatoire simple (EAS)**, chaque unité d'échantillonnage de la population a une chance égale d'être incluse dans l'échantillon. Par conséquent, chaque échantillon possible a aussi une chance égale d'être sélectionné. Pour mettre cette technique en œuvre, il faut d'abord dresser une liste de toutes les unités de la population observée.

Exemple n° 1

Pour prélever un échantillon aléatoire simple d'un annuaire téléphonique, il faudrait numéroter en ordre séquentiel chaque entrée ou inscription. S'il y avait 10 000 entrées dans l'annuaire téléphonique et si la taille de l'échantillon était de 2 000 numéros, un ordinateur devrait alors générer au hasard 2 000 numéros entre 1 et 10 000. Tous les numéros auraient la même chance d'être générés par l'ordinateur. Les 2 000 entrées de l'annuaire téléphonique correspondant aux 2 000 numéros aléatoires générés par l'ordinateur composeraient l'échantillon.

Un EAS peut être effectué avec ou sans remplacement. Un EAS avec remplacement signifierait qu'il est possible que l'entrée échantillonnée dans l'annuaire téléphonique soit sélectionnée deux fois ou plus. Habituellement, l'EAS est effectué sans remplacement parce qu'il est plus pratique et donne des résultats plus précis. Dans le reste du texte, le terme EAS sera utilisé pour faire référence à l'EAS sans remplacement, à moins d'indication contraire.

L'EAS est la méthode d'échantillonnage la plus couramment utilisée. L'avantage de cette technique tient au fait qu'elle n'exige pas d'autres données dans la base de sondage que la liste complète des membres de la population observée et l'information pour les contacter. De plus, puisque l'EAS est une méthode simple et que la théorie qui la sous-tend est bien établie, il existe des formules types pour déterminer la taille de l'échantillon, les estimations, etc., et ces formules sont faciles à utiliser.

Cependant, l'EAS nécessite une liste de toutes les unités de la population. Si cette liste n'existe pas déjà, il peut être trop dispendieux ou même irréaliste d'en créer une pour de grandes populations. Si une base de sondage est disponible et que cette base contient des informations auxiliaires, l'EAS ne permet pas de tirer parti de ces informations qui peuvent rendre d'autres méthodes d'échantillonnage plus efficaces (comme l'échantillonnage stratifié par exemple). Si la collecte doit être réalisée en personne, l'EAS pourrait donner un échantillon trop dispersé géographiquement qui ferait grimper les coûts de collecte et la durée de l'enquête.

Exemple n° 2

Imaginez que vous êtes propriétaire d'un cinéma et que vous planifiez y organiser un festival de films d'horreur le mois prochain. Pour déterminer quels films d'horreur vous y présenterez, vous voulez demander à des cinéphiles quels films ils préfèrent parmi les films que vous leur énumérerez. Pour dresser la liste des films nécessaire à votre sondage, vous décidez d'échantillonner 10 des 100 meilleurs films d'horreur de tous les temps. L'une des façons d'obtenir un échantillon consisterait à écrire tous les titres des films sur des bouts de papier, à les placer dans une boîte et à tirer les 10 titres qui constitueront votre échantillon. En utilisant cette méthode, vous auriez l'assurance que chaque film avait une probabilité égale d'être sélectionné. Vous pourriez même calculer cette probabilité, en divisant la taille de l'échantillon (n=10) par la taille de la population des 100 meilleurs films d'horreur de tous les temps (N=100). Cette probabilité serait de 0,10 (10/100), soit une chance sur dix.

Échantillonnage systématique

L'**échantillonnage systématique** signifie qu'il existe un écart, ou un intervalle, entre chaque unité sélectionnée dans l'échantillon. Par exemple, vous pourriez suivre les étapes suivantes :

1. Numérotez de 1 à **N** les unités incluses dans votre base de sondage (où **N** est la taille de la population totale).
2. Déterminez l'intervalle d'échantillonnage (**K**) en divisant le nombre d'unités dans la base de sondage par la taille de l'échantillon que vous désirez obtenir. Par exemple, pour sélectionner un échantillon de 100 unités à partir d'une population de 400, vous auriez besoin d'un intervalle d'échantillonnage de 400/100 = 4. Donc K=4. Vous devrez sélectionner une unité sur quatre pour avoir au total 100 unités à l'intérieur de votre échantillon.
3. Sélectionnez au hasard un nombre entre 1 et **K**. Ce nombre s'appelle l'origine choisie au hasard et ce sera le premier nombre inclus dans votre échantillon. Si vous choisissiez 3, la troisième unité incluse dans votre base de sondage serait la première unité comprise dans votre échantillon; si vous choisissiez 2, le début de votre échantillon serait la deuxième unité incluse dans votre base de sondage.
4. Sélectionnez chaque **K**^{e} (dans cet exemple, chaque 4e) unité après ce premier nombre. L'échantillon pourrait, par exemple, se composer des unités suivantes de façon à constituer un échantillon de 100 : 3 (l'origine choisie au hasard), 7, 11, 15, 19… 395, 399 (jusqu'à **N**, qui est 400 dans ce cas).

Vous pouvez constater que dans l'exemple ci-dessus seulement quatre échantillons sont possibles, soit ceux qui correspondent aux quatre origines possibles :

1, 5, 9, 13… 393, 397

2, 6, 10, 14… 394, 398

3, 7, 11, 15… 395, 399

4, 8, 12, 16… 396, 400

Chaque unité de la population ne fait partie que d'un seul des quatre échantillons et chaque échantillon a une probabilité égale d'être sélectionné. Chaque unité a donc une chance sur quatre de faire partie de l'échantillon, soit la même probabilité que si un EAS de taille 100 avait été tiré. La principale différence tient au fait que dans le cas d'un EAS, n'importe quelle combinaison de 100 unités aurait une chance de constituer l'échantillon, tandis que dans celui d'un échantillonnage systématique, il n'y a que quatre échantillons possibles. L'ordre des unités dans la base de sondage déterminera les échantillons systématiques possibles. Si la population est distribuée au hasard dans la base de sondage, un échantillonnage systématique devrait produire des résultats similaires à ceux d'un échantillonnage aléatoire simple.

Cette méthode est souvent utilisée dans l'industrie, où l'on sélectionne une unité pour des essais dans une chaîne de production afin de s'assurer que la machinerie et l'équipement sont d'une qualité uniforme. Un testeur dans une usine pourrait, par exemple, soumettre à un contrôle de la qualité chaque 20e produit sur une chaîne de montage, en commençant par un point initial choisi au hasard entre 1 et 20.

Les intervieweurs peuvent mettre en œuvre cette technique d'échantillonnage lorsqu'ils interrogent des gens pour une enquête-échantillon. Le responsable d'une étude de marché pourrait sélectionner, par exemple, chaque 10e personne qui entrerait dans un commerce, après avoir sélectionné au hasard la première personne. Un enquêteur pourrait interviewer les occupants de chaque 5e maison d'une rue, après avoir sélectionné au hasard l'une des cinq premières maisons.

Les avantages de l'échantillonnage systématique tiennent au fait que la sélection de l'échantillon ne peut être plus facile : vous n'obtenez qu'un seul nombre aléatoire, l'origine choisie au hasard, et le reste de l'échantillon suit automatiquement. Le plus gros inconvénient de la méthode tient au fait que les échantillons possibles risquent de ne pas être représentatifs de la population s'il existe un trait périodique dans l'ordre d'apparition des unités sur la base de sondage et que ce trait périodique coïncide d'une quelconque façon avec l'intervalle d'échantillonnage. C'est ce que l'on peut constater dans l'exemple qui suit :

Exemple n° 3

Supposez que vous dirigez une épicerie de grande surface et que vous possédez une liste des employés de chacune de ses sections. L'épicerie est divisée entre les 10 sections suivantes : le comptoir de charcuterie, la boulangerie, les caisses, les stocks, le comptoir des viandes, les fruits et légumes, la pharmacie, le magasin de photographie, le magasin de fleurs et le nettoyage à sec. Chaque section compte 10 employés, incluant un gérant (ce qui fait 100 employés au total). Votre liste est ordonnée par section, le gérant y étant énuméré le premier et les autres employés y étant ensuite inscrits dans l'ordre décroissant d'ancienneté.

Si vous voulez sonder vos employés au sujet de leurs opinions sur leur milieu de travail, vous pourriez choisir un petit échantillon pour répondre à vos questions. Si vous utilisiez un échantillonnage systématique, et si votre intervalle d'échantillonnage était 10, vous pourriez alors ne sélectionner que les gérants ou que les employés de chaque section ayant le moins d'ancienneté. Ce type d'échantillon ne vous donnerait pas un portrait complet ni approprié de l'opinion de vos employés.

Échantillonnage avec probabilité proportionnelle à la taille

Dans le contexte d'un échantillonnage probabiliste, il faut que chaque unité de la population observée ait une probabilité connue d'être incluse dans l'échantillon, mais il n'est pas nécessaire qu'elle soit la même pour tous. Si la base de sondage renferme de l'information sur la taille de chaque unité (comme le nombre d'employés de chacune des entreprises qui y sont inscrites) et si la taille de ces unités varie, on peut utiliser cette information dans le cadre de la sélection de l'échantillon afin d'en accroître l'efficacité. Cela s'appelle l'échantillonnage avec **probabilité proportionnelle à la taille**. Dans le cas de cette méthode, plus la taille de l'unité est grande, plus sa chance d'être incluse dans l'échantillon est élevée. Il faut que la mesure de la taille soit exacte pour que cette méthode augmente l'efficacité. C'est une méthode d'échantillonnage plus complexe qui ne sera pas traitée plus en détail ici.

Échantillonnage stratifié

Lorsque l'on utilise l'**échantillonnage stratifié**, on divise la population en groupes homogènes appelés strates qui sont mutuellement exclusifs, puis on sélectionne dans chaque strate des échantillons indépendants. N'importe laquelle des méthodes d'échantillonnage mentionnées dans la présente section peut être utilisée pour sélectionner l'échantillon à l'intérieur de chaque strate. La méthode d'échantillonnage peut être différente d'une strate à une autre. Toute variable pour laquelle on dispose d'une valeur pour la totalité des unités incluses dans la base de sondage (comme l'âge, le sexe, la province de résidence, le revenu, etc.) peut être utilisée pour mettre en œuvre la stratification.

Pourquoi créer des strates? Pour bien des raisons, la principale étant que leur utilisation peut rendre la stratégie d'échantillonnage plus efficace. Il a été mentionné à la section précédente que pour obtenir une estimation d'une certaine précision, il faut un échantillon plus grand pour une caractéristique qui varie beaucoup d'une unité à l'autre comparativement à une caractéristique pour laquelle la variabilité est moins grande. Si chaque personne incluse dans une population avait le même salaire, il suffirait alors d'un échantillon d'une seule unité pour obtenir une estimation précise du salaire moyen des membres de cette population.

C'est le principe qui sous-tend le gain d'efficacité réalisé grâce à la stratification. Si vous créez des strates à l'intérieur desquelles les unités auraient des caractéristiques similaires et qui différeraient considérablement de celles des unités incluses dans d'autres strates, vous n'auriez alors besoin que d'un petit échantillon tiré de chaque strate afin d'obtenir une estimation précise du revenu total pour la strate en question. Vous pourriez ensuite combiner ces estimations afin d'obtenir une estimation précise du revenu total de l'ensemble de la population. Si vous deviez utiliser un EAS de la population entière, il vous faudrait un échantillon plus grand que l'ensemble des échantillons de toutes les strates afin d'obtenir une estimation du même degré de précision pour le revenu total.

Un autre avantage est que l'échantillonnage stratifié assure d'obtenir une taille d'échantillon suffisante pour des sous-groupes d'intérêt de la population. Étant donné que chaque strate devient une population indépendante, une taille d'échantillon est déterminée pour chacune d'entre elles.

Exemple n° 4

Supposons que vous voulez estimer combien d'élèves des écoles secondaires ont un emploi à temps partiel, et ce, tant au niveau national qu'au niveau provincial. Si vous deviez sélectionner un échantillon aléatoire simple de 25 000 personnes à partir d'une liste de tous les élèves des écoles secondaires du Canada (en supposant que vous disposiez d'une telle liste), vous n'obtiendriez probablement qu'un peu plus de 100 personnes de l'Île-du-Prince-Édouard dans l'échantillon, puisque cette province représente moins 0,5 % de la population canadienne. Cet échantillon ne serait pas assez important pour le genre d'analyse détaillée que vous planifiez. Le fait de stratifier votre liste par province puis de déterminer la taille d'échantillon exacte qu'il vous faudrait pour chacune des provinces vous permettrait d'obtenir la précision souhaitée pour l'Île-du-Prince-Édouard et pour chacune des autres provinces.

La stratification est très utile lorsque les variables de stratification sont :

- simples à utiliser,
- faciles à observer,
- étroitement reliées au thème de l'enquête.

Échantillonnage par grappes

Il est parfois trop dispendieux d'avoir un échantillon réparti sur l'ensemble du territoire. Les coûts de déplacement risquent de devenir élevés si les intervieweurs doivent sonder des gens d'un bout à l'autre du pays. Les statisticiens peuvent choisir la technique de l'échantillonnage par grappes pour réduire les coûts.

La technique de l'échantillonnage par grappes implique la division de la population en groupes ou en grappes, comme son nom l'indique. Suivant cette technique, un certain nombre de grappes est sélectionné au hasard, puis toutes les unités incluses à l'intérieur des grappes sélectionnées constituent l'échantillon. Aucune unité des grappes non sélectionnées ne fait partie de l'échantillon. Elles sont représentées par les unités des grappes sélectionnées. Rappelons que dans un échantillon stratifié, des unités sont sélectionnées dans toutes les strates. C'est donc l'une des différences entre les deux approches. Entre autres exemples de grappes qui peuvent être utilisées, il y a les usines, les établissements d'enseignement et les régions géographiques telles que les subdivisions électorales.

Exemple n° 5

Supposons que vous représentez une organisation d'athlétisme désirant déterminer quels sports pratiquent les élèves de secondaire 4 (ou 11e année) au Canada. Il serait trop dispendieux et trop long d'interroger chaque élève canadien de secondaire 4 ou même deux ou trois élèves de chaque classe. Vous pourriez plutôt sélectionner au hasard 100 écoles dans tout le pays. Ces 100 écoles seraient les grappes échantillonnées. Tous les élèves de secondaire 4 de chacune des 100 écoles pourraient alors être interrogés.

L'échantillonnage par grappes crée des « poches » d'unités échantillonnées, au lieu de répartir l'échantillon également sur tout le territoire, ce qui peut permettre de réduire les coûts des opérations de collecte. Le fait de ne pas disposer d'une liste de toutes les unités incluses dans la population, mais que la liste de toutes les grappes soit disponible ou facile à dresser constitue une raison supplémentaire d'utiliser l'échantillonnage par grappes.

Dans la plupart des cas, l'échantillonnage par grappes est moins efficace qu'un EAS. C'est le principal inconvénient de cette technique. Par conséquent, il est préférable de sonder un grand nombre de petites grappes, plutôt qu'un petit nombre de grandes grappes. Pourquoi? Parce que les unités avoisinantes tendent à se ressembler davantage, ce qui donne un échantillon ne représentant pas l'éventail complet d'opinions ou de situations de l'ensemble de la population. Dans l'exemple 5, les élèves de la même école auront tendance à pratiquer les mêmes types de sports, c'est-à-dire ceux pour lesquels leur établissement d'enseignement dispose de l'équipement nécessaire.

L'échantillonnage par grappes ne permet pas de contrôler totalement la taille finale de l'échantillon, ce qui constitue un autre inconvénient de son utilisation. Puisque les écoles ne comptent pas toutes le même nombre d'élèves de secondaire 4, il se pourrait que le nombre total d'élèves de secondaire 4 dans toutes les grappes sélectionnées soit inférieur ou supérieur à la taille d'échantillon à laquelle vous vous attendiez.

Échantillonnage à plusieurs degrés

La méthode d'échantillonnage à plusieurs degrés ressemble à la méthode d'échantillonnage par grappes, sauf qu'un échantillon est prélevé à l'intérieur de chaque grappe sélectionnée. Il y a alors au moins deux degrés. Identification et sélection des grappes au premier degré, suivi d'une sélection des unités au deuxième degré à l'aide de n'importe quelle autre méthode d'échantillonnage. Dans ce contexte, les grappes sont parfois désignées comme les unités primaires d'échantillonnage (UPE) et les unités de la population comme les unités secondaires d'échantillonnage (USE). Lorsque plus de deux degrés sont utilisés, une sélection supplémentaire d'unités tertiaires d'échantillonnage (UTE) est réalisée à l'intérieur des USE, et ainsi de suite jusqu'à l'obtention d'un échantillon final.

Exemple n° 6

Dans l'exemple n° 5, 100 écoles avaient été sélectionnées au hasard et tous les élèves de secondaire 4 de ces écoles devaient être interrogés. Vous pourriez plutôt décider de sélectionner davantage d'écoles, de vous procurer une liste de tous les élèves de secondaire 4 des écoles sélectionnées et de choisir au hasard un EAS d'élèves dans chaque école. Ce serait là un plan d'échantillonnage à deux degrés. Les écoles seraient les UPE et les élèves les USE.

Vous pourriez tout aussi bien obtenir une liste de toutes les classes de secondaire 4 des écoles sélectionnées, prélever un EAS des classes de secondaire 4 dans chacune de ces écoles, vous procurer une liste de tous les élèves des classes sélectionnées et finalement choisir un échantillon d'élèves de chaque classe sélectionnée. Ce serait un plan d'échantillonnage à trois degrés. Les écoles seraient les UPE, les classes les USE et les élèves les UTE. Le processus se complique chaque fois qu'un degré est ajouté.

Imaginons maintenant que chaque école compte en moyenne 80 élèves de secondaire 4. L'échantillonnage en grappes vous permettrait d'obtenir un échantillon d'environ 8 000 élèves (100 écoles x 80 élèves). Pour avoir un échantillon de plus grande taille, vous pourriez sélectionner des écoles comptant davantage d'élèves, et pour un échantillon de plus petite taille, vous pourriez sélectionner des écoles comptant moins d'élèves. Le moyen de contrôler la taille de l'échantillon consisterait à stratifier les écoles en fonction de la taille (petite, moyenne ou grande, en référence au nombre d'élèves de secondaire 4) et à sélectionner un échantillon d'écoles dans chaque strate. On appelle cette méthode la méthode d'échantillonnage en grappes stratifiées.

Une approche alternative pour contrôler la taille de l'échantillon serait d'utiliser un plan d'échantillonnage à trois degrés. Vous pourriez sélectionner un échantillon de 400 écoles, puis sélectionner deux classes de secondaire 4 par école et finalement sélectionner 10 élèves par classe. De cette façon, vous finiriez quand même par avoir un échantillon d'environ 8 000 élèves (400 écoles x 2 classes x 10 élèves), mais l'échantillon serait davantage dispersé sur le territoire.

L'échantillonnage à plusieurs degrés permet d'obtenir un échantillon moins dispersé sur le territoire qu'avec l'EAS, par exemple, ce qui peut réduire les coûts de la collecte. Cependant, il n'est pas aussi concentré qu'avec un échantillonnage par grappes et la taille de l'échantillon nécessaire pour obtenir une certaine précision sera plus grande qu'avec l'EAS, car il est moins efficace. Il épargne quand même beaucoup de temps et d'efforts comparativement à l'EAS, parce qu'il ne nécessite pas la création d'une liste de toutes les unités de la population. Vous n'auriez pas besoin de la liste de tous les étudiants de secondaire 4 du pays, mais plutôt d'une liste des classes des 400 écoles et des élèves des 800 classes sélectionnées.

Échantillonnage à plusieurs phases

L'**échantillonnage à plusieurs phases** fait référence à la collecte de données de base auprès d'un large échantillon d'unités de la population, suivi d'une collecte de données plus détaillées pour un sous-échantillon de

ces unités. La forme la plus courante d'échantillonnage à plusieurs phases est l'échantillonnage à deux phases (ou l'échantillonnage double), mais il est également possible d'effectuer un échantillonnage à trois phases ou plus.

L'échantillonnage à plusieurs phases est assez différent de l'échantillonnage à plusieurs degrés, malgré la similarité de leurs noms. Même si l'échantillonnage à plusieurs phases implique le prélèvement de deux échantillons ou plus, la différence est que ces échantillons sont tirés de la même base de sondage. La sélection d'une unité dans la deuxième phase est conditionnelle à sa sélection dans la première phase. Une unité qui n'a pas été sélectionnée dans la première phase ne se retrouvera pas dans la seconde phase non plus. Comme dans le cas de l'échantillonnage à plusieurs degrés, plus le nombre de phases est élevé, plus le plan d'échantillonnage et l'estimation sont complexes.

L'échantillonnage à plusieurs phases est utile lorsque les informations auxiliaires qui pourraient servir à stratifier la population ou à exclure de la sélection une partie de la population ne sont pas présentes dans la base de sondage.

Exemple n° 7

Supposons qu'une organisation a besoin d'information sur des éleveurs de bétail de l'Alberta, mais que la base de sondage contient tous les types d'exploitations agricoles : d'élevage de bétail et de production laitière, de grains, de porcs, de volailles et de fruits et légumes. Pour compliquer les choses, la base de sondage ne fournit aucune donnée auxiliaire sur les exploitations agricoles qui y sont énumérées.

Il serait possible de mener une enquête toute simple dont la seule question serait : « Votre exploitation agricole est-elle en partie ou en totalité consacrée à l'élevage du bétail? » Comme elle ne compterait qu'une seule question, cette enquête devrait entraîner un faible coût par entrevue (surtout si elle est faite au téléphone), ce qui, par conséquent, permettrait à l'organisation de prélever un grand échantillon. Une fois ce premier échantillon prélevé, il serait possible d'en obtenir un second, plus petit, à partir des éleveurs de bétail, et de contacter ces exploitations agricoles pour poser des questions plus détaillées. Cette méthode éviterait à l'organisation de dépenser de l'argent pour sonder des unités ne faisant pas partie du champ d'observation (c'est-à-dire les producteurs agricoles autres que les éleveurs de bétail).

Dans l'exemple 7, l'échantillon de la première phase a été utilisé pour exclure des unités ne faisant pas partie de la population cible. Dans un autre contexte, l'information aurait pu être utilisée pour réaliser un échantillonnage plus efficace à la seconde phase, par exemple en utilisant l'information recueillie à la première phrase pour stratifier l'échantillon de la seconde phase. La méthode peut également être utilisée pour réduire le fardeau de réponse ou lorsque les coûts de collecte sont très différents d'une question de l'enquête à l'autre, comme dans l'exemple suivant.

Exemple n° 8

On pose aux participants d'une enquête sur la santé des questions de fond au sujet de leur régime alimentaire, de leur consommation de tabac et d'alcool et de leur pratique d'activité physique. Cette enquête demande en outre aux répondants de se soumettre à certains examens médicaux, comme courir sur un tapis roulant ou faire mesurer leur tension artérielle et leur taux de cholestérol.

Interroger des participants ou leur faire remplir des questionnaires sont des procédures relativement peu dispendieuses, mais les examens médicaux exigent la supervision et l'aide d'un professionnel de la santé qualifié, de même que l'utilisation d'un laboratoire équipé, ce qui peut être assez dispendieux. La meilleure façon de mener l'enquête susmentionnée consisterait à utiliser une méthode d'échantillonnage à deux phases. À la première phase, on interrogerait un échantillon d'une taille appropriée. On prélèverait à partir de cet échantillon un second échantillon plus petit. Ce sont les membres de ce second échantillon qui passeraient alors des examens médicaux.

3.2.3 Échantillonnage non probabiliste

L'échantillonnage non probabiliste est une méthode qui consiste à sélectionner des unités dans une population en utilisant une méthode subjective (c'est-à-dire non aléatoire). Comme l'échantillonnage non probabiliste ne nécessite pas de base de sondage complète, c'est un moyen rapide, facile et peu coûteux d'obtenir des données. Cependant, pour pouvoir tirer des conclusions sur la population à partir de l'échantillon, il faut supposer que l'échantillon est représentatif de la population. Il s'agit souvent d'une hypothèse risquée dans le cas d'un échantillonnage non probabiliste, car il est difficile d'évaluer si l'hypothèse est valable ou non. De plus, comme les éléments sont choisis arbitrairement, il n'y a aucun moyen d'estimer la probabilité qu'un élément soit inclus dans l'échantillon. De même, rien ne garantit que chaque élément a une chance d'être inclus, ce qui rend impossible l'estimation de la variabilité de l'échantillonnage ou l'identification d'un éventuel biais.

En général, les agences de statistiques officielles du monde entier ont utilisé l'échantillonnage probabiliste comme outil privilégié pour répondre aux besoins d'information sur une population d'intérêt. Ces dernières années, cependant, des recherches et des études ont été menées sur la manière d'appliquer l'échantillonnage non probabiliste aux statistiques officielles. L'utilisation d'autres sources de données est de plus en plus explorée. Cinq raisons principales expliquent cette tendance :

- le déclin des taux de réponse des enquêtes probabilistes,
- le coût élevé de la collecte de données,
- la charge accrue pour les répondants,
- le désir d'accéder à des statistiques en temps réel, et
- l'essor des sources de données non probabilistes telles que les enquêtes en ligne et les médias sociaux.

Certains évoquent la possibilité d'une évolution dans le paradigme et l'approche traditionnelle des statistiques. Toutefois, les données provenant de sources non probabilistes présentent quelques difficultés en ce qui a trait à la qualité des données, notamment la présence potentielle de biais de participation et de sélection. Par conséquent, les données collectées à l'aide d'un échantillonnage non probabiliste doivent être utilisées avec une prudence accrue.

Les méthodes d'échantillonnage non probabilistes couramment utilisées sont les suivantes.

Échantillonnage de commodité

Les unités sont sélectionnées de manière arbitraire, avec peu ou pas de planification. L'échantillonnage de commodité présume que les unités de la population sont toutes semblables, et que n'importe quelle unité peut être choisie pour l'échantillon. Un exemple d'échantillonnage de commodité est l'enquête de type vox pop, où l'enquêteur sélectionne une personne qu'il croise dans la rue. Malheureusement, à moins que les unités de population ne soient vraiment similaires, la sélection est sujette aux biais de l'enquêteur et de quiconque passe par là au moment de l'échantillonnage.

Échantillonnage à participation volontaire

Dans cette méthode, les répondants sont uniquement des volontaires. En général, les volontaires doivent faire l'objet d'un examen pour obtenir un ensemble de caractéristiques adaptées aux objectifs de l'enquête (par exemple, des personnes atteintes d'une maladie particulière). Cette méthode peut être sujette à d'importants biais de sélection, mais elle est parfois nécessaire. Par exemple, pour des raisons éthiques, il peut être nécessaire de solliciter des volontaires présentant des conditions médicales particulières pour certaines expériences médicales.

Voici un autre exemple d'échantillonnage à participation volontaire : au cours d'une émission radio ou télédiffusée, une question fait l'objet d'une discussion et les citoyens à l'écoute sont invités à téléphoner pour exprimer leurs opinions. Seules les personnes qui se sentent suffisamment concernées par le sujet, dans un sens ou dans l'autre, ont tendance à répondre. La majorité silencieuse ne répond généralement pas, ce qui entraîne un biais de sélection important. L'échantillonnage à participation volontaire est souvent utilisé pour sélectionner des individus pour des groupes de discussion ou des entrevues approfondies (c'est-à-dire une mise à l'essai qualitative, où l'on ne tente pas de généraliser à la population complète).

Échantillonnage au jugé

Avec cette méthode, l'échantillonnage est fait en tenant compte des idées préalables sur la composition et le comportement de la population. Un expert ayant une connaissance de la population décide quelles unités de la population doivent être choisies. En d'autres termes, l'expert sélectionne délibérément ce qu'il considère comme un échantillon représentatif. L'échantillonnage au jugé est soumis aux biais du chercheur et est peut-être encore plus biaisé que l'échantillonnage de commodité.

Puisque toutes les idées préconçues du chercheur se reflètent dans l'échantillon, des biais importants peuvent être intégrés si ces idées préconçues sont inexactes. Cependant, il peut être utile dans les études exploratoires, par exemple pour sélectionner des membres de groupes de discussion ou pour mener des entrevues approfondies afin de tester des aspects spécifiques d'un questionnaire.

Échantillonnage par quotas

Il s'agit de l'une des formes les plus courantes d'échantillonnage non probabiliste. L'échantillonnage est effectué jusqu'à ce qu'un nombre déterminé d'unités (quotas) pour diverses sous-populations soient sélectionnées. L'échantillonnage par quotas est un moyen de satisfaire les objectifs de taille d'échantillon pour les sous-populations.

Les quotas peuvent être basés sur les proportions de la population. Par exemple, si la population compte 100 hommes et 100 femmes, et il faut tirer un échantillon de 20 personnes, 10 hommes et 10 femmes peuvent être interviewés. L'échantillonnage par quotas peut être considéré comme préférable à d'autres formes d'échantillonnage non probabiliste (par exemple, l'échantillonnage au jugé), car il oblige à inclure des membres de sous-populations différentes.

L'échantillonnage par quotas ressemble quelque peu à l'échantillonnage stratifié, qui est un échantillonnage probabiliste, en ce sens que des unités similaires sont regroupées. Cependant, il diffère par la façon dont les unités sont sélectionnées. Dans l'échantillonnage probabiliste, les unités sont sélectionnées de manière aléatoire, tandis que dans l'échantillonnage par quotas, une méthode non aléatoire est utilisée. Il revient généralement à l'enquêteur de décider qui est sélectionné. Les unités contactées qui ne sont pas disposées à participer sont simplement remplacées par d'autres qui le sont, ce qui permet d'ignorer le biais de non-réponse. Les études de marché utilisent souvent l'échantillonnage par quotas (en particulier pour les enquêtes téléphoniques) au lieu de l'échantillonnage stratifié pour faire enquête auprès de citoyens ayant des profils socio-économiques particuliers. En effet, comparé à l'échantillonnage stratifié, l'échantillonnage par quotas est relativement peu coûteux, facile à administrer et présente la propriété souhaitable de satisfaire les proportions de la population. Cependant, il dissimule un biais de sélection potentiellement important.

Comme pour tous les autres plans d'échantillonnage non probabilistes, pour formuler des inférences sur la population, il faut présumer que les personnes sélectionnées sont similaires à celles qui ne le sont pas. Des hypothèses aussi fortes sont rarement valables.

Échantillonnage boule de neige ou de réseaux

Supposons qu'un chercheur souhaite trouver des individus possédant un trait rare dans la population, qu'il connaisse déjà l'existence de certains d'entre eux et sache comment les contacter. Une approche consiste à contacter ces personnes et à leur demander simplement si elles connaissent quelqu'un comme elles, puis à contacter ces personnes, etc. L'échantillon se développe comme une boule de neige dévalant une colline pour inclure, on l'espère, pratiquement toutes les personnes ayant cette caractéristique. L'échantillonnage boule de neige est utile pour les populations rares ou difficiles à atteindre, comme les personnes handicapées, les sans-abri, les toxicomanes ou d'autres personnes qui n'appartiennent pas à un groupe organisé comme les musiciens, les peintres ou les poètes, qui ne sont pas facilement identifiables sur une base de sondage. Cependant, certains individus ou sous-groupes peuvent n'avoir aucune chance d'être sélectionnés. Afin de pouvoir généraliser la conclusion à l'ensemble de la population, certaines hypothèses, qui ne sont généralement pas satisfaites, sont nécessaires.

Approche participative

L'approche participative a été définie de manière légèrement différente par les chercheurs de différents domaines. Malgré la multiplicité des définitions de l'approche participative, une constante est la communication d'un problème au public, suivi d'un appel ouvert à des contributions pour aider à résoudre le problème. Les membres du public soumettent des solutions qui appartiennent ensuite à l'entité (par exemple, des individus, des entreprises ou des organisations) qui a initialement soumis le problème. L'approche participative consiste à canaliser le désir des experts de résoudre un problème, puis à partager librement la réponse avec tout le monde.

Dans le cadre de la modernisation de Statistique Canada, l'approche participative est devenue un moyen novateur de recueillir des renseignements précieux à des fins statistiques dans certains contextes. En utilisant l'approche participative comme méthode de collecte, les enquêtes peuvent être exécutées rapidement avec un coût et un fardeau de réponse réduits. Afin de mieux comprendre les défis associés à cette approche et d'explorer la qualité des résultats, des méthodes sont développées pour comparer et valider les données à partir d'autres sources de données complémentaires. Quelques exemples sont présentés ci-dessous.

- Le projet pilote OpenStreetMap (OSM), qui s'est achevé en mars 2018, a permis de recueillir des informations géographiques grâce à l'approche participative en cartographiant les empreintes de bâtiments dans les régions d'Ottawa (Ontario) et de Gatineau (Québec). Le réseau et l'expérience de ce projet pilote ont contribué au lancement de l'initiative Bâtir le Canada 2020 (BC2020), qui vise à cartographier toutes les empreintes de bâtiments du Canada sur OSM d'ici 2020.

Panel web

Un panel web (ou panel en ligne ou internet) peut être défini comme un panel de personnes prêtes à répondre à des questionnaires en ligne. Il contient un échantillon de répondants potentiels qui ont déclaré vouloir coopérer pour une future collecte de données s'ils sont sélectionnés. Une enquête par panel web est une enquête utilisant des échantillons provenant de panels web.

Les panels web sont en quelque sorte des bases de sondage pour les enquêtes par panel web. Toutes les personnes faisant partie des panels doivent avoir une adresse courriel à jour. Le recrutement pour les panels web peut se faire de différentes manières. Les répondants peuvent être recrutés par des canaux hors ligne : téléphone, publicités télévisées, publicités radiophoniques, publicités dans les journaux et les magazines, lettres adressées, affiches extérieures, registres de clients, etc. Les répondants peuvent également provenir de canaux en ligne : courriers électroniques, sites web, bannières, sites communautaires, programmes de membres, etc. Souvent, de nombreux canaux sont utilisés afin d'obtenir la diversité nécessaire. Après le recrutement, une enquête de profil est menée afin de recueillir des informations sur les nouveaux participants au panel. Le recrutement peut se faire par le biais de panels probabilistes ou d'autorecrutement. En pratique, la distinction entre les deux peut ne pas être très importante si le taux de non-réponse est très élevé pour les panels probabilistes. Parfois, des incitations, telles que des cartes-cadeaux ou des souvenirs, sont utilisées pour attirer les gens et augmenter le taux de réponse. Les panels web sont souvent utilisés pour des recherches en marketing ou des études pilotes.

Pendant la pandémie de COVID-19, Statistique Canada a mis au point une nouvelle enquête par panel web, la Série d'enquêtes sur les perspectives canadiennes (SEPC), afin d'obtenir des renseignements en temps opportun sur la façon dont les Canadiens font face à la pandémie. Plus de 4 600 personnes dans les 10 provinces ont répondu à cette enquête entre le 29 mars et le 3 avril. Contrairement à la majorité des panels web, la SEPC est un panel probabiliste basé sur l'échantillon de l'Enquête sur la population active (EPA), certains répondants ayant accepté de répondre à de courts questionnaires en ligne à la suite de leur participation à l'EPA. La SEPC permet à Statistique Canada de recueillir des renseignements importants auprès des Canadiens de façon plus efficace, plus rapide et à moindre coût, comparativement aux méthodes d'enquête traditionnelles.

Avantages et inconvénients de l'échantillonnage non probabiliste

Avantages (+)

Rapide et pratique

En règle générale, les échantillons non probabilistes peuvent être constitués rapidement, ce qui permet de lancer, exécuter et terminer l'enquête dans des délais plus courts.

Abordable

La réalisation d'une telle enquête ne prend généralement que quelques heures à un intervieweur. De plus, comme les échantillons non probabilistes ne sont généralement pas dispersés géographiquement, les frais de déplacement des enquêteurs sont donc faibles. Dans le cas des panels web ou de l'approche participative, aucun intervieweur n'est nécessaire et le suivi des non-répondants est non requis ou moins exigeant.

Réduit le fardeau de réponse

Dans le cas de l'échantillonnage à participation volontaire et de l'approche participative, les répondants se portent eux-mêmes volontaires pour participer aux enquêtes sans avoir été sollicités personnellement.

Inconvénients (-)

Biais de sélection

Afin de faire des inférences sur la population, il est nécessaire de faire des hypothèses fortes sur la similarité entre l'échantillon et la population, même si les répondants sont autosélectionnés. En raison du biais de sélection présent dans tous les échantillons non probabilistes, il est souvent dangereux de faire ces hypothèses. Lorsqu'il s'agit de généraliser à l'ensemble de la population, il est préférable de recourir à un échantillonnage probabiliste.

Biais de non-couverture (sous-couverture)

Comme certaines unités de la population peuvent n'avoir aucune chance d'être incluses dans l'échantillon, il en résulte un biais de non-couverture. Par exemple, les personnes qui n'ont pas internet à la maison ne seront sans doute jamais sélectionnées pour un panel web et elles peuvent être différentes de celles qui ont internet.

Difficulté d'évaluation de la qualité

Il est impossible de déterminer la probabilité qu'une unité de la population soit sélectionnée pour l'échantillon, de sorte que des estimations fiables et des estimations de l'erreur d'échantillonnage ne peuvent être calculées.

3.3 Collecte

Dans la section sur l'échantillonnage, nous avons vu les différentes méthodes qui peuvent être utilisées pour sélectionner une partie de la population dans le but de mener une enquête-échantillon. Dans cette section, nous verrons les différentes méthodes pour collecter les données auprès de l'échantillon. Les éléments importants de la collecte sont le choix du mode de collecte, la conception du questionnaire et le rôle des intervieweurs. Ils sont tout aussi importants que le choix de la stratégie d'échantillonnage. D'une part à cause de leur impact sur le budget et la durée de l'enquête, mais surtout à cause de leur impact sur la qualité des données. Une stratégie de collecte appropriée devrait viser à réduire la non-réponse autant que possible tandis qu'un bon questionnaire devrait être fait de façon à minimiser le risque d'erreur de mesure. La non-réponse et l'erreur de mesure sont les sources principales d'erreur non due à l'échantillonnage.

3.3.1 Modes de collecte

Les principaux modes de collecte sont l'entrevue en personne, l'entrevue téléphonique et le questionnaire à compléter soi-même. Les entrevues peuvent être assistées par ordinateur ou non. Le questionnaire à compléter soi-même peut être sur un support papier ou numérique. Pour choisir la meilleure méthode, des facteurs comme les caractéristiques de la base de sondage, la population cible, le budget, le niveau d'exactitude requis, la sensibilité de l'information à collecter ou la complexité des concepts de l'enquête doivent être pris en considération. Voici un aperçu des différentes méthodes.

Entrevue en personne

Des intervieweurs rendent visite aux personnes dans leur milieu pour les sonder. C'est un bon moyen pour obtenir des taux de réponse élevés à un sondage ou à un recensement et, grâce au travail des intervieweurs, les données recueillies sont de meilleure qualité. Toutefois, les coûts de déplacement des intervieweurs peuvent être considérables. Si l'entrevue se déroule au domicile et qu'aucun membre du ménage n'est présent ou disponible pour répondre à l'enquête, l'intervieweur pourrait devoir repasser à plusieurs reprises pour réussir à contacter le répondant.

Lorsque l'entrevue est assistée par ordinateur, l'intervieweur apporte un ordinateur portable ou une tablette électronique et il saisit les réponses directement dans une base de données ou à l'aide d'un questionnaire électronique conçu à cet effet. Cette méthode permet de réduire le temps requis pour le traitement des données et elle évite à l'intervieweur de transporter des dizaines de questionnaires papier qui devront être rangés et protégés pour garantir la confidentialité des répondants. Elle est toutefois plus dispendieuse et plus longue à mettre en œuvre, car il faut fournir l'équipement informatique aux intervieweurs et développer et tester les systèmes avant le début de la collecte des données. Lorsque l'entrevue n'est pas assistée par ordinateur, l'intervieweur note les réponses sur des questionnaires papier. Cette méthode nécessite moins de ressources et de préparation avant la collecte, mais elle allonge le traitement des données, car les réponses doivent d'abord être saisies dans un format exploitable par l'ordinateur pour pouvoir être traitées et analysées à l'aide de logiciels.

Entrevue téléphonique

Des intervieweurs téléphonent aux personnes pour les sonder. Ce mode de collecte est plus rapide et moins dispendieux que l'entrevue en personne. Il faut toutefois disposer d'une base de sondage contenant les numéros de téléphone et la population observée exclut les ménages qui n'ont pas le téléphone. Par ailleurs, il est facile pour la personne appelée de mettre fin à l'appel si elle ne souhaite pas répondre à l'enquête… ou tout simplement de ne pas répondre au téléphone si elle ne connaît pas le numéro entrant. Par conséquent, le taux de réponse est moins élevé avec l'entrevue téléphonique qu'avec l'entrevue en personne. Il est également important de noter que de moins en moins de ménages ont des lignes téléphones fixes ce qui est susceptible d'augmenter la difficulté à les rejoindre dans le futur.

Comme pour l'entrevue en personne, l'entrevue téléphonique peut être assistée par ordinateur. Ce mode de collecte nécessite plus de préparation, mais permet de réduire le temps nécessaire pour le traitement des données.

Le questionnaire à compléter soi-même

Un questionnaire est fourni au répondant qui doit le remplir et le retourner. Il peut s'agir d'un formulaire imprimé ou d'un hyperlien et d'un code d'accès sécurisé pour le remplir en ligne. Dans les deux cas, il peut être remis à la personne sélectionnée en main propre, par la poste ou être livré à domicile. L'envoi par la poste nécessite que les adresses soient présentes dans la base de sondage. La livraison à domicile est utile lorsque les logements ne sont pas listés dans la base de sondage.

C'est le moyen le plus abordable et il peut être utilisé pour rejoindre un très grand nombre de personnes. Le répondant peut répondre aux questions au moment qui lui convient. Le questionnaire à compléter soi-même peut faciliter la collecte d'information lorsque les questions de l'enquête sont de nature délicate. Les désavantages de ce mode sont que les taux de réponse sont beaucoup plus bas que pour les autres modes de collecte et que la qualité des données collectées est moins bonne. C'est également la méthode la plus lente, car il n'est pas possible de contrôler à quel moment le répondant remplira et retournera le questionnaire d'enquête.

Le questionnaire doit être simple et accompagné d'instructions très claires, car le répondant ne peut pas demander des précisions s'il a mal compris une question. Ceux qui possèdent des capacités limitées quant à la lecture et l'écriture du français ou de l'anglais éprouveront peut-être de la difficulté à répondre au questionnaire.

Lorsqu'un support numérique est utilisé, souvent appelé le Questionnaire électronique (QE), l'interface web peut être plus conviviale pour le répondant et peut lui permettre de répondre à l'enquête plus rapidement que sur support papier. Il est possible d'indiquer au répondant s'il a omis de répondre à une question et de le renseigner sur sa progression dans le questionnaire électronique. L'envoi du questionnaire complété est presque instantané

et donc le répondant n'a pas à se rappeler de renvoyer son questionnaire par la poste comme c'est le cas avec un support papier. Comme pour les entrevues assistées par ordinateur, le temps de traitement des données est réduit, car les données sont saisies par le répondant lui-même. Cependant, les personnes qui n'ont pas accès à internet ou qui sont moins à l'aise avec les technologies numériques pourraient être moins enclines à répondre à l'enquête, à moins qu'on leur donne également l'option de répondre à un questionnaire imprimé.

De nos jours, le QE est utilisé comme première option même avant les entrevues téléphoniques par plusieurs enquêtes, incluant les enquêtes auprès des entreprises et les enquêtes sur les ménages.

Autres méthodes

D'autres méthodes utilisées pour collecter les données sont l'observation directe ou la mesure directe. En voici deux exemples. Dans le cadre de la collecte de données pour l'Indice des prix à la consommation, des enquêteurs visitent différents points de vente pour relever les prix d'une liste d'articles prédéterminée. Ce sont ces données qui permettent de mesure le taux d'inflation au Canada. Dans l'Enquête canadienne sur les mesures de la santé, certaines mesures physiques sont prises chez les répondants comme le poids, la tension artérielle ou diverses mesures d'analyse de sang. Ces mesures sont prises dans un centre d'examen mobile.

Combinaison de méthodes

La stratégie de collecte la plus efficace fera souvent appel à une combinaison des méthodes qui viennent d'être décrites. Le Recensement de la population du Canada est un bon exemple d'une stratégie de collecte multimodes. Au cours du Recensement de 2016, une lettre était envoyée par la poste à la majorité des ménages et celle-ci contenait un hyperlien et un code d'accès sécurisé pour remplir le formulaire en ligne ainsi qu'un numéro sans frais pour commander un formulaire imprimé. Dans les régions où les logements ne sont pas tous listés dans la base de sondage, des formulaires imprimés étaient livrés au domicile par des agents recenseurs responsables du listage des logements. Chaque formulaire était accompagné d'un hyperlien et d'un code d'accès sécurisé pour répondre en ligne. Dans les réserves autochtones et les régions difficiles d'accès, comme le Nunavut par exemple, la collecte était réalisée par entrevue en personne. Finalement, l'entrevue en personne et l'entrevue téléphonique ont été utilisées pour le suivi des cas pour la non-réponse. La même stratégie a été adoptée au recensement de 2021.

Une stratégie de communication utilisée de concert avec la stratégie de collecte peut être utile pour faire connaître l'enquête au public et ainsi améliorer les taux de réponse. Il peut s'agir d'une lettre pour établir un premier contact avec les ménages avant le début de la collecte par entrevues téléphoniques, d'un communiqué de presse dans les médiaux locaux ou plus rarement, d'une stratégie nationale comme dans le cas du Recensement de la population.

Qu'est-ce que les paradonnées?

Les paradonnées sont des données qui sont recueillies au cours de la collecte, mais qui portent sur l'opération de collecte en tant que telle et non sur le sujet de l'enquête. Il peut s'agir de l'heure à laquelle l'entrevue a été réalisée, la durée de l'entrevue, le nombre de tentatives pour rejoindre le répondant, etc. Elles peuvent être utilisées dans la conception de plans de collecte adaptatifs ou comme variables auxiliaires lors d'ajustements pour la non-réponse.

3.3.2 Conception du questionnaire

Le questionnaire joue un rôle de premier plan dans la démarche de collecte de données. Un questionnaire bien conçu permet de recueillir des données de manière efficace et en minimisant le risque d'erreur. Un bon questionnaire facilite le codage et la saisie des données et permet de réduire les frais et les délais de collecte et de traitement des données. Le plus grand défi dans l'élaboration d'un questionnaire consiste à traduire les objectifs de l'enquête en un cadre d'étude solide du point de vue conceptuel et méthodologique.

Avant de concevoir le questionnaire, il faut planifier l'enquête dans son ensemble, y compris les objectifs, les besoins en données et l'analyse. Une fois le questionnaire élaboré, il faudra en faire l'essai avant de procéder à la collecte des données.

Une foule de choses est à considérer lorsque l'on entreprend de concevoir un questionnaire. Voici des considérations qui entrent en jeu :

- L'introduction est-elle informative? Suscite-t-elle l'intérêt des répondants?
- Les termes employés sont-ils simples, directs et familiers à l'ensemble des répondants?
- Les questions se lisent-elles bien? L'ensemble du questionnaire est-il cohérent?
- Les questions sont-elles claires et précises?
- Le questionnaire commence-t-il par des questions faciles et intéressantes?
- A-t-on précisé la période de référence?
- Y trouve-t-on de doubles questions?
- Y a-t-il des questions tendancieuses?
- Les questions devraient-elles être dirigées ou non dirigées? Si elles sont dirigées, est-ce que les catégories de réponses sont mutuellement exclusives et collectivement exhaustives?
- Les questions s'appliquent-elles à tous les répondants?

Introduction et conclusion du questionnaire

L'introduction du questionnaire est très importante, car elle présente l'information pertinente de l'enquête. Ainsi, l'introduction devrait :

- indiquer le titre ou l'objet de l'enquête,
- nommer l'organisme qui fait l'enquête,
- présenter l'objectif de l'enquête,
- demander la collaboration des répondants,
- informer les répondants à propos des dispositions en matière de confidentialité,
- préciser le caractère obligatoire ou volontaire de l'enquête, et
- déclarer s'il existe des ententes de partage des données avec d'autres organisations.

Les répondants se demandent souvent si l'information recueillie sera utile. Par conséquent, on doit souligner l'importance de remplir le questionnaire et préciser la manière dont l'information sera utilisée et dont les répondants pourront y avoir accès. Lorsqu'on mène une enquête, il est essentiel que les répondants puissent bien comprendre quelle sera la valeur de l'information.

Voici un exemple d'une introduction efficace :

Évaluation des besoins des élèves

Nom de l'école _____

Veuillez prendre le temps nécessaire pour remplir le présent questionnaire (environ 50 à 75 minutes). Grâce à vos réponses, votre école aura accès à des renseignements importants qui lui permettront d'adopter des solutions efficaces pour améliorer votre santé et votre bien-être.

Renseignements confidentiels

Quel est le but de cette enquête?

Cette enquête vous donne l'occasion de partager vos idées quant aux façons d'assurer votre bien-être et votre sécurité dans l'école.

Vous n'êtes pas tenu de remplir ce questionnaire. Toutefois, l'opinion de chaque élève est importante et si les élèves participent en grand nombre à cette enquête, les résultats seront plus précis. Soyez assuré que le présent questionnaire est entièrement confidentiel.

1. N'écrivez pas votre nom sur le questionnaire.
2. Veuillez sceller votre questionnaire dans l'enveloppe ci-jointe.

Par la suite, l'enveloppe sera ouverte par les membres de l'équipe qui effectueront la saisie par ordinateur des réponses que vous aurez fournies. Votre enveloppe sera ajoutée à d'autres; ainsi, il sera impossible d'identifier les répondants. Les résultats de **tous** les questionnaires seront compilés et remis à la direction de votre école.

Les premières questions d'une enquête devraient encourager les répondants à avoir confiance en leur capacité de répondre aux questions suivantes. Si nécessaire, les premières questions devraient aider à identifier le répondant en tant que membre de la population visée par l'enquête.

La conclusion d'un questionnaire bien conçu contient une section de commentaires pour permettre au répondant d'inscrire des éléments qui ne figurent pas dans le questionnaire. De cette façon, on évitera toute frustration de la part du répondant et on lui donnera l'occasion d'exprimer ses idées, questions ou inquiétudes. Finalement, on devrait remercier les répondants d'avoir pris le temps de remplir le questionnaire.

Formulation des questions

Dans toute enquête, la formulation des questions demeure un enjeu fondamental. Les questions et les consignes doivent être faciles à comprendre et la manière de les formuler représente un élément considérable. La même question reformulée peut apporter des résultats tout à fait différents. Ainsi, on peut songer aux aspects suivants.

Abréviations et acronymes

Remplacez les abréviations et acronymes par la formulation complète.

Exemple : Savez-vous si les chiffres de la pop sont offerts en direct?

On pourrait mieux formuler la question de la façon suivante : Savez-vous si les chiffres de la population du Recensement de 2006 sont présentés dans le site Internet www.statcan.gc.ca?

Exemple : Avez-vous déjà participé à l'EPA?

On pourrait mieux formuler la question de la façon suivante : Avez-vous déjà participé à l'Enquête sur la population active de Statistique Canada?

Formulation et terminologie complexes

Évitez d'utiliser une terminologie spécialisée ou complexe.

Exemple : Savez-vous qui mène les discussions sur la fusion imminente des localités voisines en nouvelles régions métropolitaines?

On pourrait mieux formuler la question de la façon suivante : Savez-vous qui mène les discussions dans chacune des provinces sur la fusion de villes, de villages et de zones rurales en nouvelles régions métropolitaines?

Exemple : Le vaccin antipneumococcique vous a-t-il été administré?

On pourrait mieux formuler la question de la façon suivante : Avez-vous été vacciné contre la grippe?

Cadre de référence

Donnez tous les détails concernant le cadre de référence de la question.

Exemple : « Quel est votre revenu? »

Le mot « votre » vise-t-il le revenu du répondant, le revenu familial ou le revenu du ménage? Par « revenu », entend-on le salaire seulement, les pourboires ou le revenu d'autres sources? Puisqu'on indique aucune période précise, est-ce qu'il s'agit du revenu de la dernière semaine, du dernier mois ou de la dernière année?

Cette question est trop vague; on doit la reformuler en ajoutant des détails précis selon le cadre de référence.

On pourrait mieux formuler la question de la façon suivante : Quel a été le revenu total de toute source de votre ménage avant impôt et déductions l'an dernier?

Questions précises

Le cadre de référence d'une question n'est pas le seul élément qui requiert des détails précis. Afin d'obtenir des réponses uniformes dans l'ensemble de l'échantillon, les questions doivent parfois énoncer le genre de réponses à donner.

Exemple : On montre une bouteille de jus d'orange au répondant et on lui demande : Quelle quantité de jus d'orange y a-t-il dans cette bouteille?

Voici des réponses qu'on peut obtenir :

- Une orange, un peu d'eau et du sucre
- De l'orange à 25 % et de l'eau gazéifiée à 75 %
- Du jus d'une demi-douzaine d'oranges
- Trois onces de jus d'orange
- Du jus entier
- Un quart de tasse de jus d'orange
- Pas de jus d'orange
- Pas beaucoup de jus d'orange
- Ne sais pas
- Une chopine
- La majeure partie du contenu
- Environ un verre et demi

On pourrait mieux formuler la question de la façon suivante : Cette bouteille contient 250 ml d'une boisson à l'orange. Combien de millilitres de jus d'orange contient-elle?

Doubles questions

Exemples :

Planifiez-vous de laisser votre voiture à la maison et de prendre l'autobus pour aller au travail au cours de l'année à venir?

Votre entreprise offre-t-elle de la formation pour les nouveaux employés et du recyclage pour le personnel en place?

Dans ces exemples, on se trouve en réalité à poser deux questions plutôt qu'une.

Dans le premier exemple, on demande aux répondants s'ils pensent laisser leur voiture à la maison et s'ils comptent prendre l'autobus l'année prochaine.

Dans le deuxième exemple, on demande aux répondants si leur entreprise offre de la formation aux nouveaux employés et si elle offre du recyclage pour le personnel en place.

Parfois, la réponse sera la même aux deux parties de la question. Toutefois, on pourrait aussi donner deux réponses bien différentes à chaque partie de la question. Le répondant aurait donc de la difficulté à interpréter la question.

La meilleure solution serait de formuler deux questions séparées.

Questions tendancieuses

L'exemple suivant illustre l'effet que produit une question tendancieuse :

Exemple 1 :

À votre avis, devrait-on pouvoir faire des achats le dimanche en Ontario; en d'autres termes, les magasins qui veulent rester ouverts le dimanche devraient-ils pouvoir le faire?

- Résultats :
  - 73 % sont en faveur du magasinage le dimanche
  - 25 % sont contre le magasinage le dimanche
  - 2 % n'ont pas d'opinion

Exemple 2 :

À votre avis, le dimanche devrait-il être un jour chômé en Ontario; en d'autres termes, le gouvernement devrait-il faire du dimanche le seul jour de la semaine où la plupart des gens n'ont pas à travailler?

- Résultats :
  - 50 % sont contre un dimanche chômé
  - 44 % sont en faveur d'un dimanche chômé
  - 6 % sont sans opinion

Source : Enquête de 1991 dans la région métropolitaine de Toronto.

Dans la première question, on demande aux répondants s'ils sont en faveur du magasinage le dimanche, tandis que dans la deuxième, on leur demande s'ils sont d'accord pour ne pas travailler le dimanche. Par conséquent, les résultats étaient très différents.

À propos de la différence des résultats, on pourrait donner une explication selon laquelle certains répondants n'avaient pas bien compris les implications dans cette question. Certaines personnes peuvent se prononcer contre le travail le dimanche, mais sont en faveur du magasinage. Toutefois, si personne ne travaille le dimanche, les commerces ne peuvent être ouverts pour les consommateurs!

Questions ouvertes ou fermées

En général, il y a deux genres de questions : **ouvertes** ou **fermées**. Aux questions ouvertes, on répond dans ses propres mots et aux questions fermées, on répond en cochant une case ou en encerclant la bonne réponse.

Question ouverte

Quelle est la plus importante question à laquelle font face les jeunes aujourd'hui?

Question fermée

Parmi les questions auxquelles font face les jeunes aujourd'hui, laquelle est la plus importante?

- Chômage
- Unité nationale
- Environnement
- Violence chez les jeunes
- Hausse des frais de scolarité
- Drogue dans les écoles
- Nécessité de disposer d'un plus grand nombre d'ordinateurs dans les écoles
- Orientation professionnelle

Les deux genres de questions présentent des avantages et des inconvénients. Une question ouverte permet au répondant d'interpréter ce qu'on lui dit et de répondre comme il le veut. Celui-ci donne sa réponse par écrit ou l'intervieweur inscrit mot à mot la réponse qu'il donne.

La question fermée oblige le répondant à choisir une réponse à partir des options de réponses. Pour le répondant, il est plus facile et plus rapide de répondre à des questions fermées. Dans le cas d'un chercheur, il est plus facile et moins coûteux de coder et d'analyser les réponses. En outre, les questions fermées produisent des réponses plus cohérentes, ce qui n'est pas toujours le cas dans les questions ouvertes.

Essai du questionnaire

Il s'agit d'une étape fondamentale dans l'élaboration d'un questionnaire permet :

- de découvrir des anomalies dans la formulation ou l'ordre des questions,
- de trouver les erreurs dans la présentation ou les instructions d'un questionnaire,
- de cerner les problèmes occasionnés par l'incapacité ou le refus de répondre aux questions,
- de proposer des catégories de réponses complémentaires qui peuvent être précodées dans le questionnaire,
- de fournir une indication provisoire de la durée de l'entrevue et des cas de refus.

On peut effectuer un essai du questionnaire complet ou d'une seule portion du questionnaire. À un moment donné, on devra cependant faire un essai complet du questionnaire.

3.3.3 Rôle des intervieweurs

Il importe de noter que les gens qui recueillent des données ne sont pas tous des intervieweurs. Parfois, des personnes se rendent dans les épiceries et les magasins de vêtements afin de recueillir des données. Elles relèvent manuellement les prix d'une liste de biens et de services pour ensuite communiquer ces données au personnel de Statistique Canada.

Il reste que l'intervieweur joue un rôle fort important. Pour interroger des gens aux fins de collecte de données, il faut un certain nombre de compétences, sans quoi cela pourrait nuire à la qualité des données obtenues. Ainsi, les compétences recherchées chez un intervieweur sont les suivantes :

- posséder une bonne aptitude à communiquer,
- avoir une personnalité empreinte d'assurance et de professionnalisme, et
- posséder un permis de conduire, dans le cas des intervieweurs qui devront se déplacer pour procéder à des entrevues en personne.

Statistique Canada affecte un grand nombre d'intervieweurs à la collecte des données. Avant d'entreprendre leur travail, ceux-ci reçoivent une formation où l'on met l'accent sur la présentation et la manière de se présenter aux répondants, puisque cela influera grandement sur leur réaction et leur volonté de collaborer. Ainsi, les intervieweurs doivent d'abord suivre certaines règles avant d'interroger les répondants, telles que :

- se présenter et montrer une pièce d'identité au répondant,
- expliquer qu'une enquête est en cours et préciser qui la réalise,
- exposer le but de l'enquête,
- indiquer au répondant que son ménage ou son entreprise a été sélectionné,
- donner au répondant le temps de lire ou de s'informer sur la confidentialité, sur le caractère obligatoire ou volontaire de l'enquête et sur l'échange de données avec d'autres organismes, et
- lire le message d'introduction du questionnaire au répondant.

Il importe en outre que l'intervieweur ait les compétences voulues, dont les suivantes :

- savoir susciter l'intérêt du répondant,
- savoir écouter attentivement,
- savoir poser les questions à chaque répondant telles qu'elles sont formulées,
- ne pas suggérer de réponses au répondant,
- bien répondre aux questions du répondant,
- savoir garder le répondant informé, et
- savoir exposer les règles de confidentialité des renseignements recueillis.

Par-dessus tout, l'intervieweur doit bien faire sentir au répondant qu'il écoute ce qu'il dit.

3.4 Traitement

Dans la première moitié de cette section, nous avons abordé en détail les nombreuses étapes de la collecte de données dans le cadre d'une enquête ou d'un recensement : la phase de planification, les différentes manières de sélectionner un échantillon, la façon d'atteindre les unités sélectionnées et de recueillir les informations nécessaires pour atteindre l'objectif de l'enquête à l'aide d'un questionnaire bien conçu et souvent avec l'aide d'intervieweurs. Une fois les données obtenues, soit à partir de cette série d'étapes, soit à partir d'une source d'information administrative ou alternative, il est temps de les traiter afin qu'elles soient prêtes à être utilisées pour produire des informations statistiques.

Cette section décrit les cinq étapes de traitement qui sont couramment utilisées pour transformer les données avant l'estimation. La complexité du traitement ainsi que le temps et les ressources nécessaires à la préparation des données dépendent du type de source et de la qualité des données.

3.4.1 Codage

Le codage est tout processus qui attribue une valeur (un code) à une réponse. Cela signifie que le codage consiste soit à attribuer un code à une réponse donnée, soit à comparer la réponse à un ensemble de codes et à sélectionner celui qui décrit le mieux la réponse. Le code peut être une valeur numérique ou une chaîne de

caractères. Il peut y avoir différentes manières de réaliser cette traduction, mais les différentes approches de codage affectent la qualité et le coût des données produites.

Les questionnaires comportent généralement deux types de questions : les questions fermées et les questions ouvertes. Les réponses à ces questions affectent le type de codage effectué. La question suivante est un exemple de question fermée :

Dans quelle mesure le sport est-il important pour vous procurer les avantages suivants?

<1/> Très important

<2/> Quelque peu important

<3/> Pas important

La question suivante est un exemple de question ouverte :

Quels sont les sports que vous pratiquez?

Veuillez préciser_____

Dans le cas des questions fermées, les catégories de réponse sont déterminées avant la collecte, le code apparaissant généralement sur le questionnaire à côté de chaque catégorie de réponse. Pour les questions ouvertes, le codage a lieu après la collecte et peut être manuel ou automatisé. Pour certaines questions, le codage peut être simple (par exemple, l'état civil). Dans des cas plus complexes, comme la géographie, l'industrie et la profession, un système de codage standard est fortement recommandé lorsqu'il est disponible. Mais pour de nombreuses questions pour lesquelles il n'existe pas de système de codage standard, déterminer un bon schéma de codage est une tâche non triviale.

Systèmes de codage automatisés

Le codage manuel nécessite une interprétation et un jugement de la part du codeur, et peut varier d'un codeur à l'autre. En raison des progrès technologiques, des contraintes de ressources et, surtout, des préoccupations relatives à la rapidité et à la qualité, le codage devient de plus en plus automatisé.

En général, deux fichiers sont entrés dans un système de codage automatisé. Un fichier, appelé le fichier d'entrée, contient soit les réponses à l'enquête, soit les fichiers administratifs qui doivent être codés. L'autre fichier est appelé le fichier de référence, qui contient l'ensemble de codes prédéterminé. Ensuite, pour chaque enregistrement du fichier d'entrée, une recherche est effectuée dans le fichier de référence. Si une correspondance est trouvée, le code dans le fichier de référence est attribué à l'enregistrement correspondant du fichier d'entrée. Sinon, le code est laissé en blanc. Certains des avantages d'un système de codage automatisé sont que le processus devient de plus en plus rapide, cohérent et abordable.

De nombreux systèmes automatisés sont déjà utilisés à Statistique Canada. Par exemple, les fichiers de données de l'Enquête sur la population active sont recueillis auprès des bureaux régionaux de Statistique Canada et passent par un système de codage automatisé qui attribue des codes d'industrie et de profession basés sur le Système de classification des industries de l'Amérique du Nord (SCIAN) et la Classification nationale des professions (CNP). Les enregistrements rejetés (ceux qui n'ont pas de correspondance avec la réponse écrite) sont les seules données à être codées manuellement.

Récemment, des techniques d'apprentissage automatique ont été utilisées par le Registre des entreprises de Statistique Canada pour faciliter l'attribution des codes industriels à partir des noms et adresses des entreprises. Cela permet d'améliorer la couverture du Registre des entreprises, qui est la base de sondage de la majorité des enquêtes auprès des entreprises de Statistique Canada, et, par conséquent, d'améliorer la qualité des données de nombreuses enquêtes auprès des entreprises.

3.4.2 Saisie

La saisie des données est n'importe quel processus qui permet de convertir les données dans un format exploitable par un ordinateur. Dans la vie quotidienne, une foule de dispositifs permettent de saisir les données.

Supposons que vous vous levez un matin et consultez votre téléphone intelligent pour connaître la météo. Vous utiliserez l'écran tactile de votre téléphone pour sélectionner l'application appropriée. Ensuite, pour consulter les nouvelles, vous pourriez ouvrir votre navigateur et saisir le nom de votre page de nouvelle préférée à l'aide du clavier à l'écran. Un peu plus tard, vous sortez pour faire des courses. Vous portez au poignet une montre d'entraînement qui utilise des capteurs pour relever votre nombre de pas et votre fréquence cardiaque. Rendu à la caisse, le caissier utilisera un scanneur pour relever le code-barre de chaque article qui correspond au nom de l'article et au prix correspondant dans la base de données. Si vous payez à l'aide d'une carte bancaire, le caissier vous tendra un terminal de paiement électronique. Celui-ci prélèvera l'information sur votre carte grâce à la puce qu'elle contient ou par un système de paiement sans contact. Il se pourrait qu'on vous demande de saisir votre numéro d'identification personnel (NIP) à l'aide du clavier. De retour à la maison, vous décidez de jouer à un jeu vidéo pour vous divertir. Vous utilisez une télécommande pour sélectionner le jeu de votre choix et une manette pour contrôler vos actions dans le jeu. Ou alors vous vous asseyez devant votre ordinateur portable pour répondre à vos courriels. Vous utiliserez le pavé tactile ou une souris pour ouvrir la bonne application logicielle et le clavier pour taper vos réponses.

Tous ces dispositifs vous ont permis de saisir les données nécessaires aux activités de votre journée :

- Écran tactile
- Clavier à l'écran
- Montre équipée de capteurs
- Scanneur
- Terminal de paiement électronique
- Télécommande
- Manette de jeu
- Clavier physique
- Pavé tactile
- Souris

À cela pourraient s'ajouter caméra et enregistreur vocal de votre téléphone intelligent.

Tous ces dispositifs et bien d'autres pourraient potentiellement servir à saisir des données dans le cadre d'une enquête, dépendant de l'information qui doit être collectée. Dans le cas d'une enquête qui nécessite de collecter les données auprès de répondants, un intervieweur équipé d'un ordinateur ou le répondant lui-même, si on lui fournit un questionnaire électronique à compléter soi-même, saisira les données à l'aide de la souris et du clavier. Un téléphone intelligent ou une tablette électronique pourraient également être utilisés à cet effet. D'autres moyens peuvent être utilisés pour coordonner la collecte, par exemple des systèmes de code-barre ou de NIP qui permettent de faire le lien entre la personne sélectionnée dans la base de sondage et les données saisies.

Un aspect de la saisie des données qui vous est sans doute moins familier est la manière de procéder dans le cas où les données d'enquête sont collectées grâce à un formulaire imprimé. Ce mode peut être plus pratique et rapide dans certains contextes, mais requiert plus de travail pour saisir les données.

La saisie peut être faite manuellement. Il s'agit d'une personne qui lit les réponses inscrites sur le formulaire imprimé, les code et les saisit elle-même à l'aide du clavier d'un ordinateur. Dans ce cas, un bon système de codage est essentiel pour accélérer la saisie et réduire le risque d'erreurs.

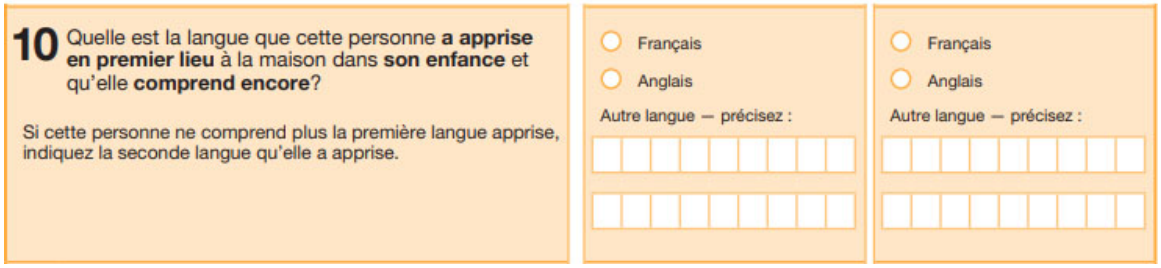
La saisie peut aussi être faite automatiquement grâce à la lecture optique des questionnaires. Dans le cas des questions fermées, la Reconnaissance optique des marques (ROM) permet de détecter les cases qui ont été cochées par un répondant. Par exemple, la figure 3.4.2.1 montre comment à la question 8 du Recensement

de 2021, « Cette personne connaît-elle assez bien le français ou l'anglais pour soutenir une conversation? », le répondant doit cocher un cercle correspondant à l'un des quatre choix de réponse proposés.

Figure 3.4.2.1
Exemple de cases à cocher pour répondre à une question fermée

Dans le cas des questions ouvertes, la Reconnaissance intelligence des caractères (RIC) permet de saisir et d'identifier les lettres inscrites dans chacune des cases. La figure 3.4.2.2 montre comment à la question 10 du Recensement de 2021, « Quelle est la langue que cette personne a apprise en premier lieu à la maison dans son enfance et qu'elle comprend encore? », le répondant doit indiquer la langue en inscrivant une seule lettre dans chacune des cases prévues à cet effet.

Figure 3.4.2.2
Exemple de cases à remplir à l'aide de lettres moulées pour répondre à une question ouverte



Qu'elle soit faite de façon manuelle ou automatique, la saisie des réponses à un formulaire imprimé peut être sujette à des erreurs. Il est donc essentiel de mettre en place des processus pour en contrôler la qualité.

3.4.3 Vérification

Dans un monde idéal, les données seraient collectées sans aucune erreur. Malheureusement, les réponses, qu'elles proviennent d'enquêtes ou de fichiers administratifs, peuvent être manquantes, incomplètes ou incorrectes. La vérification des données est l'application de contrôles pour détecter les entrées manquantes, invalides ou incohérentes ou pour indiquer les enregistrements de données qui sont potentiellement erronés. Quel que soit le type de données avec lequel vous travaillez, certaines vérifications doivent être effectuées à différentes étapes ou phases de la collecte et du traitement des données. La vérification des données est décrite et illustrée ici en se concentrant sur les enquêtes, mais elle est aussi largement appliquée à d'autres sources de données, telles que les données administratives, pour assurer la qualité des données.

La vérification des données commence par la question suivante : « Quelles pourraient être les causes des erreurs dans nos fichiers? » Il existe plusieurs situations où des erreurs peuvent se glisser dans les données, et la liste suivante en donne quelques-unes :

- Un répondant pourrait avoir mal compris une question.
- Un répondant ou un intervieweur pourrait avoir coché la mauvaise réponse.
- Un codeur pourrait avoir mal codé ou mal compris une réponse écrite.
- Un intervieweur pourrait avoir oublié de poser une question ou d'enregistrer la réponse.
- Un répondant pourrait avoir fourni des réponses inexactes.
- Certaines questions pourraient avoir été laissées en blanc.

Gardez toujours à l'esprit les objectifs de la vérification des données :

- assurer l'exactitude des données,
- établir la cohérence des données,
- déterminer si les données sont complètes,
- assurer la cohérence des données agrégées,
- obtenir les meilleures données possible.

Appliquer les règles de vérification

Alors, comment procéder à la vérification des données? La première étape consiste à appliquer des règles, ou des facteurs à prendre en considération, aux données. Ces règles sont déterminées à partir de l'expertise d'un spécialiste du sujet, de la structure du questionnaire, de l'historique des données et de toute autre enquête ou tout autre ensemble de données connexe.

Les connaissances spécialisées peuvent provenir de diverses sources. Le spécialiste peut être un analyste qui a une grande expérience du type de données à éditer. Un expert peut également être l'un des commanditaires de l'enquête qui connaît bien les relations entre les données.

La présentation et la structure du questionnaire auront également un impact sur les règles de vérification des données. Par exemple, il est parfois demandé aux répondants d'ignorer certaines questions si celles-ci ne s'appliquent pas à eux ou à leur situation. Cette spécification doit être respectée et intégrée dans les règles de vérification.

Enfin, d'autres sources de données relatives au même type de variables ou de caractéristiques sont utilisées afin d'établir certaines des règles de vérification des données. Par exemple, les enquêtes auprès des entreprises collectent généralement des données financières sur les entreprises. Les mêmes informations peuvent être disponibles dans les déclarations fiscales de l'entreprise. Ainsi, les données fiscales peuvent être utilisées pour développer des règles de vérification pour valider les données d'enquête.

Types de vérification de données

Il existe plusieurs types de vérification de données couramment utilisés, notamment :

- Les **vérifications de validité** portent sur un champ ou une cellule à la fois. Elles s'assurent que les identificateurs d'enregistrement, les caractères invalides et les valeurs ont été pris en compte, que les champs essentiels ont été remplis (par exemple, aucun champ de quantité n'est laissé blanc alors qu'un nombre est requis), que les unités de mesure spécifiées ont été correctement utilisées et que les données déclarées se situent dans l'étendue des valeurs autorisées (par exemple, l'heure de déclaration se situe dans les limites spécifiées). Dans le cadre de la collecte de données assistée par ordinateur, comme les questionnaires électroniques, la vérification des données en temps réel est généralement intégrée au système de collecte de données afin que la validité des données soit évaluée au fur et à mesure de leur collecte.
- Les **vérifications des doublons** examinent un enregistrement complet à la fois. Ces types de vérification permettent d'éviter les enregistrements en double, en s'assurant qu'un répondant ou une unité d'enquête n'a été enregistré qu'une seule fois. Une vérification des doublons permet également de s'assurer que le répondant n'apparaît pas plus d'une fois dans l'univers de l'enquête, surtout s'il y a eu un changement de nom. Enfin, il garantit que les données n'ont été saisies qu'une seule fois dans le système.
- Les **vérifications de cohérence** comparent différentes réponses d'un même enregistrement pour s'assurer qu'elles sont cohérentes entre elles. Par exemple, si une personne est déclarée comme appartenant au groupe d'âge des 0 à 14 ans, mais qu'elle déclare également être retraitée, il y a un problème de cohérence entre les deux réponses. Les vérifications interchamps sont une autre forme de vérification de la cohérence. Ces vérifications permettent de s'assurer que si un chiffre est déclaré dans une section, un chiffre correspondant est déclaré dans une autre.

- Les **vérifications historiques** sont utilisées pour comparer les réponses de l'enquête actuelle et précédente. Par exemple, tout changement radical depuis la dernière enquête sera signalé. Les ratios et les calculs sont également comparés, et tout écart de pourcentage qui sort des limites établies sera noté et remis en question.
- Les **vérifications statistiques** portent sur l'ensemble des données. Ce type de vérification n'est effectué qu'après que toutes les autres vérifications ont été appliquées et que les données ont été corrigées. Les données sont compilées et toutes les valeurs extrêmes, les données suspectes et les valeurs aberrantes sont rejetées.
- Les **vérifications diverses** comprennent les dispositions spéciales de déclaration, les vérifications dynamiques propres à l'enquête, les vérifications de classification correcte, les changements d'adresse physiques, de lieux ou de contacts, et les vérifications de lisibilité (c'est-à-dire s'assurer que les chiffres ou les symboles sont reconnaissables et faciles à lire).

La vérification des données est influencée par la complexité du questionnaire. La complexité fait référence à la longueur, ainsi qu'au nombre de questions posées. Elle comprend également le détail des questions et l'éventail des sujets que le questionnaire peut couvrir. Dans certains cas, la terminologie d'une question peut être très technique. Pour ces types d'enquêtes, il peut y avoir des arrangements spéciaux pour les rapports et des vérifications spécifiques à l'industrie.

Niveaux de vérification des données

La vérification des données peut être effectuée manuellement, avec l'aide d'un programme informatique, ou une combinaison des deux techniques. Selon le support (électronique, papier) par lequel les données sont soumises, il existe deux niveaux de vérification des données : la microvérification et la macro-vérification.

- La microvérification consiste à corriger les données au niveau de l'enregistrement. Ce processus vise à détecter les erreurs en vérifiant les enregistrements de données individuels. L'objectif à ce stade est de déterminer la cohérence des données et de corriger chaque enregistrement.
- La macro-vérification vise à détecter également les erreurs, mais elle le fait par l'analyse des données agrégées (totaux). Les données sont comparées à celles d'autres enquêtes, de fichiers administratifs ou de versions antérieures des mêmes données. Ce processus permet de déterminer la comparabilité des données.

3.4.4 Imputation

La vérification n'a que peu de valeur pour l'amélioration globale des résultats réels de l'enquête si aucune mesure corrective n'est prise lorsque les éléments ne respectent pas les règles établies au cours du processus de vérification. Lorsque toutes les données ont été vérifiées à l'aide des règles appliquées et qu'un fichier présente des données manquantes, l'imputation est généralement effectuée dans le cadre d'une étape distincte.

Les valeurs manquantes ou invalides ont un impact certain sur la qualité des résultats de l'enquête. L'imputation est le processus utilisé pour attribuer des valeurs de remplacement aux valeurs manquantes, invalides ou incohérentes qui ont échoué aux vérifications. Cette opération intervient après un suivi des répondants (si possible), une révision manuelle et une correction des questionnaires (le cas échéant). À ce stade, tous les types d'erreurs sont corrigés, y compris les erreurs commises par les répondants et les erreurs survenues lors du codage et de la saisie des données.

Les procédures d'imputation visent à combler les lacunes. En général des modifications sont apportées à un nombre minimal de champs jusqu'à ce que l'enregistrement complet passe toutes les vérifications. Lorsque ces erreurs sont détectées, les valeurs des entrées invalides, manquantes ou incomplètes sont imputées ou remplacées par des valeurs appropriées, et des réponses sont fournies pour les questions sans réponse. Cette procédure est mieux accomplie par ceux qui ont un accès complet aux microdonnées et qui sont en possession de bonnes informations auxiliaires.

Bien que l'imputation puisse améliorer la qualité des données finales, il faut veiller à choisir une méthode d'imputation appropriée. Certaines méthodes d'imputation ne préservent pas la relation entre les variables. En fait, certaines peuvent même fausser les distributions sous-jacentes.

Voici quelques méthodes d'imputation des données couramment utilisées :

- L'**imputation déductive** est généralement la première méthode utilisée. Cette méthode est utilisée lorsqu'une valeur peut être déduite avec certitude et qu'elle peut être réalisée pendant la collecte, la saisie, la vérification ou les étapes ultérieures du traitement des données. L'imputation déductive est utilisée lorsqu'il n'y a qu'une seule réponse possible à la question (par exemple, toutes les valeurs sont données, mais le total ou le sous-total est manquant).
- L'**imputation par donneur de l'enquête (hot deck)** utilise les valeurs provenant d'un autre enregistrement de la même enquête, qui est désigné comme le donneur, afin de répondre à la question (ou à la série de questions) qui nécessite une imputation. Le donneur peut être sélectionné de manière aléatoire à partir d'un groupe de donneurs présentant le même ensemble de caractéristiques prédéterminées. Par exemple, si un questionnaire a été retourné sans indication du revenu annuel, nous pouvons déterminer les caractéristiques du donneur comme étant des enregistrements ayant la même province, la même profession et le même niveau d'expérience que le répondant de l'enquête nécessitant une imputation. Une liste de donneurs possibles correspondant à ces critères est créée et l'un d'entre eux est sélectionné au hasard. Une fois le donneur trouvé, la réponse du donneur (dans ce cas, le revenu annuel) remplace la réponse manquante ou invalide.
- L'**imputation par donneur d'une autre source (cold deck)** est similaire à l'imputation par donneur de l'enquête. La différence est que cette dernière utilise des donneurs de la même enquête tandis que l'imputation par donneur d'une autre source utilise des donneurs d'une autre source, comme des données historiques d'une itération antérieure de la même enquête ou des données administratives.
- L'**imputation par valeur moyenne** consiste à remplacer la valeur manquante ou incohérente par la valeur moyenne calculée à partir des unités répondantes ayant le même ensemble de caractéristiques prédéterminées. Par exemple, s'il manque dans un enregistrement un chiffre total pour le revenu annuel d'un individu, on peut imputer le revenu moyen observé dans la province de cet individu pour la même profession avec le même niveau d'expérience que le répondant. L'un des inconvénients de l'imputation par valeur moyenne est qu'elle détruit la distribution et les relations entre les variables en créant un pic artificiel à la moyenne du groupe. Cela réduit artificiellement la variance échantillonnale estimée si l'on utilise les formules conventionnelles de la variance échantillonnale.
- L'**imputation par voisin le plus proche** est un autre type d'imputation par donneur. Dans ce cas, il faut élaborer une sorte de critère pour déterminer l'unité répondante qui ressemble le plus à l'unité ayant la valeur manquante, conformément aux caractéristiques prédéterminées. L'unité la plus proche de la valeur manquante est alors utilisée comme donneur.

Il existe d'autres méthodes d'imputation plus sophistiquées, qui utilisent la modélisation statistique pour attribuer une valeur de remplacement.

La méthode d'imputation peut varier d'une enquête à l'autre et même, dans des circonstances particulières, au sein d'une même enquête. Très souvent, différentes méthodes sont combinées entre elles afin de fournir la valeur la plus appropriée pour une variable. Ces méthodes peuvent être appliquées manuellement ou à l'aide d'un système automatisé. Pour faciliter cette tâche, Statistique Canada a mis au point un système généralisé d'imputation pour imputer les données sur la base de l'expertise de statisticiens expérimentés qui ont analysé l'enquête et suggéré les approches pour imputer des données significatives.

Bien que l'imputation puisse améliorer la qualité des données finales, il faut faire preuve de prudence dans le choix d'une méthode d'imputation appropriée. L'un des risques de l'imputation est qu'elle peut détruire des données déclarées pour créer des enregistrements correspondant à des modèles préconçus qui peuvent s'avérer incorrects par la suite. L'adéquation des méthodes d'imputation dépend de l'enquête, de ses objectifs, des informations auxiliaires disponibles et de la nature de l'erreur.

En outre, toutes les méthodes d'imputation peuvent être appliquées à d'autres sources de données, sans se limiter aux données d'enquête. Par exemple, Statistique Canada reçoit et utilise des données financières de

l'Agence du revenu du Canada afin de réduire le fardeau de réponse, et ces données administratives comportent souvent des valeurs manquantes ou incohérentes. Afin d'en faire bon usage, des systèmes rigoureux de vérification et d'imputation ont été mis en place pour améliorer la qualité des données avant de passer à l'étape suivante.

Notez également que dans le cas de la non-réponse totale, lorsque très peu ou pas de données ont été collectées pour un enregistrement ou une unité, une approche courante consiste à effectuer une repondération pour tenir compte de la non-réponse, une technique dont il sera question dans la section sur l'estimation.

3.4.5 Couplage d'enregistrements

Le couplage d'enregistrements est le processus par lequel des enregistrements ou des unités provenant de différentes sources de données sont réunis dans un seul fichier à l'aide d'identifiants non uniques, tels que des noms, des dates de naissance, des adresses et d'autres caractéristiques. Il est également connu sous le nom d'appariement de données, de couplage de données, de résolution d'entités et de nombreux autres termes selon les domaines dans lesquels il a été utilisé. L'idée initiale du couplage d'enregistrements remonte aux années 1950, puis cette technique a été appliquée par des personnes issues d'un large éventail de domaines, tels que l'entreposage de données et l'intelligence de gestion, la recherche historique, ainsi que la pratique et la recherche médicales.

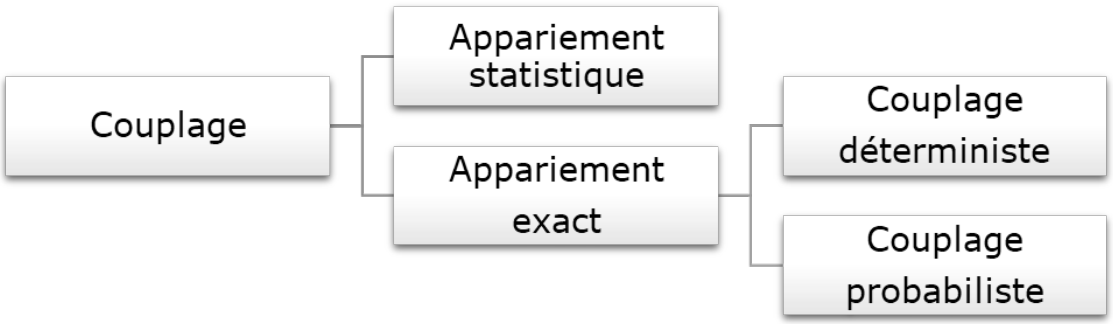
Le couplage a une longue histoire d'utilisations dans les enquêtes statistiques et le développement de données administratives. À Statistique Canada, le couplage d'enregistrements est utilisé pour créer une base de sondage, éliminer les doublons des fichiers, fournir des renseignements supplémentaires pour faciliter le traitement des données ou combiner des fichiers de façon à étudier les relations entre deux ou plusieurs éléments de données provenant de fichiers distincts. Par exemple :

- Un registre des entreprises comprenant des noms, des adresses et d'autres informations d'identification, telles que des informations financières complètes, peut être construit à partir de bases de données sur les impôts et l'emploi.
- Une enquête sur les établissements de vente au détail ou les établissements agricoles pourrait combiner les résultats d'une base aréolaire et d'une base liste. Pour produire un estimateur pour la combinaison des sources, les unités de la base aréolaire devraient être identifiées dans la liste.
- La couverture du Recensement de la population peut être mesurée en couplant les enregistrements du Recensement à d'autres sources de données administratives et en estimant le pourcentage de personnes trouvées dans l'une des sources, mais pas dans l'autre.

Types de couplage

Il existe deux types de couplage d'enregistrements : l'appariement exact et l'appariement statistique. L'appariement statistique se divise en deux sous-types : le couplage d'enregistrements déterministe et le couplage d'enregistrements probabiliste, tel qu'illustré à la figure 3.4.5.1 ci-dessous.

Figure 3.4.5.1
Types de couplage d'enregistrements



Appariement statistique

L'objectif de l'appariement statistique est de créer un fichier reflétant la distribution de la population sous-jacente. Les enregistrements qui sont combinés ne correspondent pas nécessairement à la même entité, telle qu'une personne ou une entreprise. Les fichiers qui sont appariés peuvent avoir des unités différentes, mais se référer à la même population. On suppose que la relation des variables dans la population sera similaire à la relation dans les fichiers. Cette méthode est principalement utilisée dans les études de marché et rarement par les agences statistiques officielles.

Appariement exact

L'objectif de l'appariement exact est de relier les informations relatives à un enregistrement particulier dans un fichier aux informations d'un fichier secondaire afin de créer un seul fichier avec des informations correctes pour chaque enregistrement. Le couplage est effectué au niveau de l'enregistrement, par exemple un lien entre les enregistrements de mortalité et le recensement de la population.

Couplage d'enregistrements déterministe

Il s'agit de la forme la plus simple de couplage d'enregistrements, qui produit des liens basés sur des identifiants ou des variables communes parmi les sources de données disponibles. Il arrive souvent qu'il n'existe pas de variable unique exempte d'erreurs, présente sur la majorité des données et ayant un pouvoir de discrimination suffisant. Seule une combinaison de variables sera capable de discriminer entre deux enregistrements. C'est une technique souvent utilisée par les agences de statistiques officielles. Statistique Canada utilise cette méthode pour construire ses registres d'entreprises, d'adresses et de population, ce qui implique de multiples opérations d'enquête par la suite.

Couplage d'enregistrements probabiliste

Il s'agit d'un autre type d'appariement exact. Comme dans l'autre cas, il n'y a pas d'identifiant unique disponible pour l'appariement. Contrairement à l'appariement déterministe, l'appariement probabiliste peut compenser si les informations sont incomplètes ou sujettes à erreur. Les enregistrements qui ne concordent pas totalement pour chaque variable peuvent être reliés entre eux pour constituer un ensemble de paires potentielles. Un score est alors calculé pour chaque paire potentielle. Ensuite, un statut de couplage est attribué à chaque paire potentielle sur la base du score.

Remarque

De nombreux facteurs sont à prendre en compte pour déterminer le type de couplage d'enregistrements à utiliser, comme l'objectif du couplage, le type de données, le coût, le temps, la confidentialité, le niveau de précision acceptable et le type d'erreur. En général, le couplage déterministe est moins exigeant sur le plan informatique, mais il implique davantage d'interventions manuelles. Le couplage probabiliste est plus long et plus intensif sur le plan informatique, et nécessite un logiciel spécialisé. Cependant, il produit des résultats généralement plus fiables que le couplage déterministe.

3.5 Estimation

Comme nous le savons maintenant, le but de la collecte de données, y compris la réalisation d'enquêtes ou l'acquisition de données à partir d'autres sources, est d'obtenir des informations sur une population particulière. Une fois l'échantillon sélectionné, les informations collectées (voir la section sur la collecte des données) et les données traitées (voir la section sur le traitement des données), il reste encore à relier les informations recueillies auprès de l'échantillon à la population cible.

En général, l'estimation consiste à utiliser la valeur dérivée d'un échantillon pour estimer la valeur d'une caractéristique correspondante de la population. Les chercheurs souhaitent généralement examiner les estimations de nombreuses statistiques pour différentes variables, les totaux, moyennes et proportions étant les plus fréquentes. Par exemple, une enquête-échantillon pourrait être utilisée pour produire l'une des statistiques suivantes :

- la proportion de fumeurs parmi toutes les personnes âgées de 15 à 24 ans dans la population,
- les revenus moyens des hommes et des femmes ayant un diplôme universitaire, et
- le nombre total de voitures possédées par l'ensemble de la population de l'enquête.

Dans cette section, nous décrirons ce qu'est le processus d'estimation de façon générale, en commençant par la pondération, suivi par l'estimation de l'erreur due à l'échantillonnage et par une description des différentes sources d'erreurs non dues à l'échantillonnage. Tout comme l'échantillonnage, l'estimation nécessite des connaissances avancées en statistique mathématique. Les exemples présentés dans cette section sont basés sur le plan d'échantillonnage le plus simple, l'échantillonnage aléatoire simple, et ne visent qu'à donner un aperçu de l'estimation.

3.5.1 Pondération

Le principe de l'estimation dans une enquête probabiliste est que chaque unité de l'échantillon représente non seulement elle-même, mais aussi plusieurs unités de la population de l'enquête. Le poids d'échantillonnage d'une unité fait habituellement référence au nombre moyen d'unités de la population que chaque unité échantillonnée représente. La détermination de ce poids découle directement du plan d'échantillonnage et est une partie importante du processus d'estimation.

Bien que les poids d'échantillonnage puissent être utilisés pour l'estimation, la plupart des enquêtes produisent un ensemble de poids d'estimation en ajustant les poids d'échantillonnage pour améliorer la précision des estimations finales. Les deux raisons les plus courantes de procéder à des ajustements sont de tenir compte de la non-réponse et d'utiliser des données pertinentes provenant d'autres sources. Une fois que les poids d'estimation finaux ont été calculés, ils sont appliqués aux données de l'échantillon afin de calculer les estimations.

Poids d'échantillonnage

La première étape de l'estimation consiste à attribuer un poids à chaque unité échantillonnée. Le **poids d'échantillonnage** (w_{ech}), qui est le nombre moyen d'unités de la population que chaque unité échantillonnée représente, est l'inverse de sa probabilité d'inclusion (π) dans l'échantillon.

$$w_{ech} = 1 / \pi$$

Si la probabilité d'inclusion est de 1/50, alors chaque unité sélectionnée représente en moyenne 50 unités dans la population et le poids d'échantillonnage est $w_{ech} = 50$.

Certains plans d'échantillonnage attribuent les mêmes poids d'échantillonnage à toutes les unités de l'échantillon, tandis que d'autres donnent des poids d'échantillonnage différents aux unités échantillonnées pour diverses raisons, comme l'amélioration de la précision ou la réduction des coûts.

Exemple 1 : Échantillonnage aléatoire simple

Supposons qu'il y ait N =100 élèves de secondaire 5 (ou 12e année) dans une école secondaire. Un échantillon aléatoire simple de taille n =25 élèves est tiré et les élèves sélectionnés sont invités à remplir un questionnaire sur leur plan de carrière.

- La probabilité d'inclusion est :
  $\pi = n / N = 25 / 100 = 1 / 4$.
- Le poids d'échantillonnage est :
  $$w_{ech} = \frac{1}{\pi} = 1 / \frac{1}{4} = 4.$$

Chaque élève sélectionné dans cet échantillon représente quatre élèves de l'école.

Production d'estimations simples

Les estimations peuvent être produites après le calcul des poids, mais seules les estimations simples, telles que les totaux, les moyennes et les proportions, sont couvertes ici.

Estimation d'un total de la population

L'estimation du nombre total ($\hat{Y}$) d'unités dans la population est calculée en multipliant le poids et la valeur d'intérêt pour chaque unité sélectionnée puis en additionnant toutes les unités de l'échantillon. Pour les variables catégoriques, l'estimation est en fait calculée en additionnant les poids des unités répondantes.

Exemple 2 : Échantillonnage aléatoire simple (suite)

Supposons que parmi les 25 élèves sélectionnés dans l'échantillon, environ 10 ont postulé à des programmes scientifiques. Alors, le nombre total d'étudiants ayant postulé à des programmes scientifiques est de :

$$\hat{Y} = 4 \times 10 = 40$$

Estimation d'une moyenne de la population

L'estimation de la moyenne ($\hat{\bar{Y}}$) dans la population est l'estimation de la valeur totale de la variable d'intérêt ($\hat{\bar{Y}}$) divisée par l'estimation du nombre total d'unités ($\hat{N}$) dans la population.

$$\hat{\bar{Y}} = \frac{\hat{Y}}{\hat{N}}$$

Exemple 3 : Échantillonnage aléatoire simple (suite)

En général, les étudiants postulent à plus d'un programme d'études postsecondaires. Supposons que parmi les 25 étudiants sélectionnés dans l'échantillon, 5 d'entre eux ne posent leur candidature qu'à un seul programme, 10 d'entre eux posent leur candidature à deux programmes et 10 d'entre eux posent leur candidature à trois programmes. Alors, le nombre moyen de candidatures par étudiant est calculé comme ci-dessous :

- Le nombre total de candidatures est donné par :

$$\hat{Y} = \left(4 \times 5 \times 1\right) + \left(4 \times 10 \times 2\right) + \left(4 \times 10 \times 3\right) = 220$$

- Le nombre total d'étudiants est donné par :

$$\hat{N} = 4 \times 25 = 100$$

- Le nombre moyen de candidatures par étudiant est donné par :

$$\hat{\bar{Y}} = \frac{\hat{Y}}{\hat{N}} = \frac{220}{100} = 2.2$$

Estimation d'une proportion de la population

L'estimation de la proportion de la population de l'enquête ayant une caractéristique donnée est assez similaire à l'estimation d'une moyenne de population en termes de formule mathématique. Elle est également calculée comme un quotient entre deux totaux estimés. La principale différence réside dans le numérateur, qui indique l'estimation du nombre total d'unités possédant la caractéristique donnée (c) lors de l'estimation d'une proportion ($\hat{P}$). En revanche, le numérateur indique l'estimation de la valeur totale pour les données quantitatives lors de l'estimation d'une moyenne.

$$\hat{P} = \frac{\widehat{Nc}}{\hat{N}}$$

Exemple 4 : Échantillonnage aléatoire simple (suite)

Supposons que parmi les 25 élèves sélectionnés dans l'échantillon, il y ait 10 femmes et 15 hommes. Au total, 10 élèves, dont 5 femmes et 5 hommes, s'inscrivent à un programme scientifique. La proportion d'élèves qui s'inscrivent à un programme scientifique par sexe est calculée comme ci-dessous :

- Le nombre total d'étudiants inscrits à un programme scientifique par sexe est donné par :

$$\hat{N}_{homme,\,science} = 5 \times 4 = 20$$

$$\hat{N}_{femme,\,science} = 5 \times 4 = 20$$

- Le nombre total d'étudiants par sexe est donné par :

$$\hat{N}_{homme} = 15 \times 4 = 60$$

$$\hat{N}_{femme} = 10 \times 4 = 40$$

- La proportion d'étudiants appliquant le programme scientifique par sexe est donnée par :

$$\hat{P}_{homme,\,science} = \frac{\hat{N}_{homme,\,science}}{\hat{N}_{homme}} = \frac{20}{60} = 1/3$$

$$\hat{P}_{femme,\,science} = \frac{\hat{N}_{femme,\,science}}{\hat{N}_{femme}} = \frac{20}{40} = 1/2$$

Autres méthodes d'estimation

La méthode d'estimation décrite ci-dessus pour l'échantillonnage aléatoire simple est la méthode d'estimation la plus simple. Il en existe d'autres, plus avancées, qui sont largement appliquées dans de nombreuses enquêtes. La méthode d'estimation la plus appropriée à utiliser est déterminée par quelques facteurs, tels que les caractéristiques à estimer, les différents types de données, la fiabilité, le coût et l'actualité, etc. À Statistique Canada, des systèmes d'estimation spécialisés sont utilisés pour produire des estimations impliquant des procédures compliquées en temps opportun.

Ajustements à la pondération

Très souvent, les poids d'échantillonnage doivent être ajustés avant l'estimation, et il y a deux types principaux d'ajustement : l'ajustement pour la non-réponse et l'ajustement pour l'information externe.

Ajustement pour la non-réponse

Presque toutes les enquêtes souffrent de non-réponse, ce qui se produit lorsque toutes ou certaines informations clés demandées aux unités échantillonnées ne sont pas disponibles pour certaines raisons, telles que le refus de participer de l'unité échantillonnée, l'absence de contact, l'impossibilité de localiser l'unité ou l'impossibilité d'utiliser les informations obtenues. La façon la plus simple de traiter une telle non-réponse est de l'ignorer, mais ceci peut conduire à des estimations inexactes.

Deux façons courantes de traiter la non-réponse sont d'imputer les réponses manquantes ou d'ajuster les poids d'échantillonnage pour que les unités répondantes représentent à la fois les unités répondantes et non répondantes. Les poids d'échantillonnage des non-répondants sont alors redistribués parmi les répondants.

Ajustement pour l'information externe

Parfois, des informations sur la population de l'enquête sont disponibles à partir d'autres sources, par exemple des informations provenant d'un recensement ou d'un fichier administratif. Ces informations peuvent également être incorporées dans le processus de pondération.

Il y a deux raisons principales pour utiliser des données externes (auxiliaires) lors de l'estimation. La première raison est qu'il est souvent important que les estimations de l'enquête correspondent à des totaux de population connus ou à des estimations provenant d'une autre enquête plus fiable. Par exemple, de nombreuses enquêtes sociales ajustent leurs estimations d'enquête afin d'être cohérentes avec les estimations (répartitions par âge, sexe, etc.) du dernier recensement de la population. Des informations externes peuvent également être obtenues à partir de données administratives ou d'une autre enquête considérée comme plus fiable en raison de la taille plus importante de son échantillon ou parce que ses estimations publiées doivent être respectées.

La deuxième raison est d'améliorer la précision des estimations, pourvu que les valeurs des variables auxiliaires soient collectées pour les unités enquêtées et que des totaux de population ou des estimations soient disponibles pour ces variables à partir d'une autre source fiable.

3.5.2 Erreur d'échantillonnage

Une autre partie importante de l'estimation consiste à estimer l'ampleur de l'erreur d'échantillonnage dans l'estimation. Ceci permet de mesurer la précision des estimations de l'enquête pour le plan d'échantillonnage spécifique. L'erreur d'échantillonnage ne peut être estimée que si un échantillonnage probabiliste est utilisé.

L'erreur d'échantillonnage est l'erreur causée par l'observation d'un échantillon au lieu de l'ensemble de la population. Elle résulte de l'estimation d'une caractéristique de la population en examinant seulement une partie de la population plutôt que la population entière, et se réfère à la différence entre l'estimation dérivée d'une enquête par sondage et la vraie valeur qui serait obtenue si un recensement de la population entière était effectué dans les mêmes conditions. Il n'y a pas d'erreur d'échantillonnage dans un recensement, car les calculs sont basés sur l'ensemble de la population.

Estimation de l'erreur d'échantillonnage

Comme mentionné précédemment, toute estimation dérivée d'un échantillon est sujette à une erreur d'échantillonnage, car seule une partie de la population a été observée. Un échantillon différent pourrait produire des estimations différentes. L'erreur d'échantillonnage est à l'origine de la variabilité des estimations dérivées de différents échantillons, même si la taille et le plan d'échantillonnage sont identiques, de même que la méthode d'estimation utilisée. Elle est généralement mesurée par la variance d'échantillonnage, qui dépend de nombreux facteurs, notamment la méthode d'échantillonnage, la méthode d'estimation, la taille de l'échantillon et la variabilité de la caractéristique estimée.

Variance d'échantillonnage

Dans les plans d'échantillonnage simples, comme l'échantillonnage aléatoire simple, la variance d'échantillonnage peut être calculée directement à l'aide d'une formule. Cependant, il n'existe généralement pas de formule pour les plans plus complexes. Dans ce cas, une estimation de la variance d'échantillonnage peut être calculée en utilisant des méthodes telles que la linéarisation de Taylor ou des méthodes de rééchantillonnage telles que le jackknife et le bootstrap.

Quelle que soit la méthode utilisée pour l'estimation de la variance, elle doit intégrer les propriétés du plan d'échantillonnage telles que la stratification, la mise en grappes et la sélection en plusieurs étapes ou phases, le cas échéant.

Les autres facteurs qui influent sur l'ampleur de la variance d'échantillonnage sont les suivants :

- En général, la variance d'échantillonnage diminue lorsque la **taille de l'échantillon** augmente, mais le changement n'est pas proportionnel.
- La **taille de la population** a un impact sur la variance d'échantillonnage pour les populations de taille petite à moyenne. Pour les grandes populations, son impact est mineur.

- La **variabilité de la caractéristique d'intérêt dans la population** affecte également la taille de l'erreur d'échantillonnage. Plus la différence entre les unités de la population est grande, plus la taille de l'échantillon nécessaire pour atteindre un niveau de précision spécifique est importante.
- Le **plan d'échantillonnage**, qui comprend un plan de sondage et une procédure d'estimation, influe également sur l'ampleur de l'erreur d'échantillonnage. La méthode d'échantillonnage, appelée « plan d'échantillonnage », peut grandement affecter la taille de l'erreur d'échantillonnage. Les enquêtes impliquant un plan d'échantillonnage complexe peuvent entraîner une erreur d'échantillonnage plus importante qu'un plan plus simple. La procédure d'estimation a également un impact majeur sur l'erreur d'échantillonnage. Ces concepts sont examinés plus en détail dans la section portant sur l'échantillonnage.

Autres mesures de l'erreur d'échantillonnage

Outre l'utilisation de la variance d'échantillonnage pour mesurer l'erreur d'échantillonnage, il existe d'autres méthodes fréquemment utilisées, notamment : l'erreur type, le coefficient de variation, la marge d'erreur et l'intervalle de confiance.

- L'**erreur type** est la racine carrée de la variance d'échantillonnage. Cette mesure est plus facile à interpréter puisqu'elle donne une indication de l'erreur d'échantillonnage en utilisant la même échelle que l'estimation alors que la variance est basée sur les différences au carré.
- Le **coefficient de variation (CV)** évalue la taille de l'erreur type par rapport à l'estimation de la caractéristique mesurée. Il s'agit du ratio entre l'erreur type de l'estimation et la valeur moyenne de l'estimation elle-même. Le CV est très utile pour comparer la précision des estimations d'un échantillon, lorsque leurs tailles ou leurs échelles diffèrent les unes des autres. Même si le CV est largement utilisé dans les publications officielles de Statistique Canada, il n'est pas recommandé pour mesurer la précision des proportions, surtout lorsque les proportions estimées sont proches de 0 ou de 1. Dans ce cas, il est plus approprié d'utiliser l'intervalle de confiance.
- L'**intervalle de confiance (IC)** donne un intervalle de valeurs autour de l'estimation qui a une certaine probabilité d'inclure la vraie valeur de la mesure d'intérêt dans la population. Cette probabilité est le niveau de confiance de l'IC. Pour une estimation donnée dans un échantillon donné, l'utilisation d'un niveau de confiance plus élevé génère un IC plus large, c'est-à-dire un IC moins précis. Le niveau de confiance le plus couramment utilisé est de 95 %, mais des niveaux de confiance de 99 % ou 90 % sont également utilisés dans certaines circonstances.
- La **marge d'erreur** correspond à la moitié de la largeur de l'IC. Plus la marge d'erreur est grande, moins on peut avoir confiance dans le fait que le résultat d'un sondage reflète le résultat d'une enquête sur l'ensemble de la population. Elle est souvent utilisée pour rendre compte de l'erreur d'échantillonnage par les sondeurs ou les journalistes.

Exemple 1

Il est fréquent de voir les résultats d'une enquête publiés dans un journal comme ci-dessous :

> Selon un récent sondage, 15 % des résidents d'Ottawa assistent à des services religieux chaque semaine. Les résultats, basés sur un échantillon de 1 345 résidents, sont considérés comme exacts à **plus ou moins trois points de pourcentage 19 fois sur 20**.

Dans cet exemple, l'expression « 19 fois sur 20 » signifie que, si le sondage était répété de nombreuses fois, alors l'intervalle de confiance couvrirait la vraie valeur dans la population 19 fois sur 20. C'est équivalent à un niveau de confiance de 95 %. L'expression « plus ou moins trois points de pourcentage » signifie que la marge d'erreur est de 3 %. Par conséquent, la valeur de l'estimation est de 15 % et l'IC à 95 % correspondant à cette estimation est de 12 % à 18 %.

3.5.3 Erreur non due à l'échantillonnage

L'**erreur non due à l'échantillonnage** réfère à toutes les sources d'erreur qui ne sont pas liées à l'échantillonnage. Les erreurs non dues à l'échantillonnage sont présentes dans tous les types d'enquête, incluant les recensements et les données administratives. Elles se produisent pour un certain nombre de raisons : la base de sondage peut être incomplète, certains répondants peuvent ne pas déclarer les données avec exactitude, les données peuvent manquer pour certains répondants, etc.

Les erreurs non dues à l'échantillonnage peuvent être classées en deux groupes : les **erreurs aléatoires** et les **erreurs systématiques**.

- **Les erreurs aléatoires** sont des erreurs dont les effets s'annulent approximativement si l'on utilise un échantillon suffisamment grand, ce qui entraîne une augmentation de la variabilité.
- **Les erreurs systématiques** sont des erreurs qui ont tendance à aller dans le même sens et donc qui s'accumulent sur l'ensemble de l'échantillon, entraînant un biais dans les résultats finaux. Contrairement aux erreurs aléatoires, ce biais n'est pas réduit par l'augmentation de la taille de l'échantillon. Les erreurs systématiques sont la principale cause d'inquiétude en ce qui concerne la qualité des données d'une enquête. Malheureusement, les erreurs non dues à l'échantillonnage sont souvent extrêmement difficiles, voire impossibles, à mesurer.

Types d'erreur non due à l'échantillonnage

Les erreurs non dues à l'échantillonnage peuvent se produire dans tous les aspects du processus d'enquête et peuvent être classées dans les catégories suivantes : erreur de couverture, erreur de mesure, erreur de non-réponse et erreur de traitement.

Erreur de couverture

L'erreur de couverture consiste en des omissions (sous-couverture), des inclusions erronées, des duplications et de mauvaises classifications (surcouverture) d'unités dans la base de sondage. Comme elles affectent chaque estimation produite par l'enquête, elles constituent l'un des types d'erreurs les plus importants. Dans le cas d'un recensement, elle peut être la principale source d'erreur. L'erreur de couverture peut avoir des dimensions à la fois spatiales et temporelles, et peut entraîner un biais dans les estimations. L'effet peut varier pour différents sous-groupes de la population. Cette erreur a tendance à être systématique et est généralement due à une sous-couverture, c'est pourquoi il est important de la réduire autant que possible.

Erreur de mesure

L'erreur de mesure, également appelée **erreur de réponse**, est la différence entre les valeurs mesurées et les vraies valeurs. Elle se compose d'un biais et d'une variance et résulte de données incorrectement demandées, fournies, reçues ou enregistrées. Ces erreurs peuvent être dues à des inefficacités du questionnaire, de l'intervieweur, du répondant ou du processus d'enquête.

- **Mauvaise conception du questionnaire**
  Il est essentiel que les questions soient formulées avec soin afin d'éviter de créer des biais. Si les questions sont trompeuses ou prêtent à confusion, les réponses peuvent être faussées.
- **Biais de l'intervieweur**
  Un intervieweur peut influencer la façon dont une personne répond aux questions de l'enquête. Cela peut se produire lorsque l'intervieweur est trop amical ou distant ou qu'il incite le répondant. Pour éviter cela, les intervieweurs doivent être formés pour rester neutres tout au long de l'interview. Ils doivent également faire très attention à la façon dont ils posent chaque question. Si l'intervieweur modifie la formulation d'une question, cela peut avoir un impact sur la réponse de la personne interrogée.
- **Erreur du répondant**
  Les répondants peuvent également fournir des réponses incorrectes. Des souvenirs erronés, des tendances à exagérer ou à minimiser les événements, et des inclinations à donner des réponses qui semblent plus socialement acceptables sont plusieurs raisons pour lesquelles une personne interrogée peut donner une fausse réponse.
- **Problèmes liés au processus d'enquête**
  Des erreurs peuvent également se produire en raison d'un problème lié au processus d'enquête lui-même. L'utilisation de réponses de substitution, c'est-à-dire des réponses obtenues d'une personne autre que le répondant, ou le manque de contrôle sur les procédures d'enquête ne sont que quelques-uns des facteurs qui augmentent le risque d'erreurs de réponse.

Erreur de non-réponse

Les estimations obtenues après l'observation d'une non-réponse et le recours à l'imputation pour traiter cette non-réponse ne sont généralement pas équivalentes aux estimations qui auraient été obtenues si toutes les valeurs souhaitées avaient été observées sans erreur. La différence entre ces deux types d'estimations s'appelle l'erreur de non-réponse. Il existe deux types d'erreurs de non-réponse : totale et partielle.

- **L'erreur de non-réponse totale** se produit lorsque toutes les réponses ou presque d'une unité d'échantillonnage sont manquantes. Ceci peut survenir si le répondant n'est pas disponible ou temporairement absent, qu'il ne peut pas participer ou qu'il refuse de participer à l'enquête, ou si le logement est vacant. Si un nombre important d'unités échantillonnées ne répondent pas à une enquête, les résultats peuvent être biaisés puisque les caractéristiques des non-répondants peuvent différer de celles des participants.
- **L'erreur de non-réponse partielle** se produit lorsque le répondant fournit des informations incomplètes. Pour certaines personnes, quelques questions peuvent être difficiles à comprendre, ou elles peuvent refuser ou oublier de répondre à une question. Un questionnaire mal conçu ou de mauvaises techniques d'entrevue peuvent également être à l'origine d'une erreur de non-réponse partielle. Pour réduire cette forme d'erreur, il convient d'apporter un soin particulier à la conception et au test des questionnaires. Une formation adéquate des intervieweurs et des stratégies de vérification et d'imputation appropriées contribueront également à minimiser cette erreur.

Erreur de traitement

L'**erreur de traitement** se produit pendant le traitement des données. Elle comprend toutes les activités de traitement des données après la collecte et avant l'estimation, telles que les erreurs de saisie, de codage, de vérification et de tabulation des données ainsi que d'affectation des poids de l'enquête.

- Les **erreurs de codage** se produisent lorsque différents codeurs codent différemment la même réponse, ce qui peut être causé par une mauvaise formation, des instructions incomplètes, une variation de la performance du codeur (c'est-à-dire la fatigue, la maladie), des erreurs de saisie des données ou un mauvais fonctionnement de la machine (certaines erreurs de traitement sont causées par des erreurs dans les programmes informatiques).
- Les **erreurs de saisie** de données se produisent lorsque les données ne sont pas saisies dans l'ordinateur exactement comme elles apparaissent sur le questionnaire. Cela peut être causé par la complexité des données alphanumériques et par le manque de clarté de la réponse fournie. La disposition physique du

questionnaire lui-même ou des documents de codage peut provoquer des erreurs de saisie des données. La méthode de saisie des données, manuelle ou automatisée (par exemple, à l'aide d'un lecteur optique), peut également entraîner des erreurs.

- Les **erreurs de vérification et d'imputation** peuvent être causées par la mauvaise qualité des données d'origine ou par leur structure complexe. Lorsque les processus de vérification et d'imputation sont automatisés, les erreurs peuvent également résulter de programmes défectueux insuffisamment testés. Le choix d'une méthode d'imputation inappropriée peut entraîner un biais. Les erreurs peuvent également résulter de la modification incorrecte de données qui se sont avérées erronées, ou de la modification par erreur de données correctes.

3.6 Gestion de la qualité

La qualité est un facteur essentiel à tous les niveaux du traitement. La réputation de Statistique Canada à titre de meilleur organisme statistique du monde repose sur la qualité de ses données. Pour assurer la qualité d'un produit ou d'un service lors des activités d'élaboration d'une enquête, nous devons recourir à des méthodes d'**assurance de la qualité** et de **contrôle de la qualité**.

Assurance de la qualité

L'assurance de la qualité désigne toutes les activités prévues destinées à inspirer la confiance en laquelle un produit ou un service répondra aux objectifs et aux besoins des utilisateurs. Dans le contexte des activités d'enquête, l'assurance de la qualité peut avoir lieu à chacune des principales étapes de l'élaboration d'une enquête : la planification, la conception, la mise en œuvre, le traitement, l'évaluation et la diffusion.

Voici des exemples d'activités prévues :

- améliorer la base de sondage d'une enquête,
- modifier la conception de l'échantillon,
- modifier le processus de collecte des données,
- améliorer les tâches régulières de suivi,
- modifier les procédures de traitement, et
- réviser la conception du questionnaire.

L'assurance de la qualité vise à rehausser la qualité en prévoyant les problèmes avant qu'ils ne surviennent et à assurer la qualité par l'utilisation des techniques de prévention et de contrôle.

Contrôle de la qualité

Le contrôle de la qualité est une procédure réglementaire qui permet :

- la mesure de la qualité,
- la comparaison de la qualité avec des normes établies, et
- la réaction aux différences entre les valeurs mesurées et les normes établies.

Par exemple, notons le contrôle de la qualité des activités de collecte par entrevue, de codage et de saisie des données.

Le contrôle de la qualité a pour but d'atteindre un certain niveau de qualité à un coût minimal. Certaines fonctions d'assurance et de contrôle sont souvent accomplies au sein d'une unité d'enquête, surtout lors du codage des données, de la saisie et de la vérification. Certaines procédures sont automatisées, d'autres sont partiellement automatisées et d'autres encore sont entièrement manuelles.

Différences entre l'assurance qualité et le contrôle de la qualité

Assurance de la qualité

- prévoit les problèmes avant qu'ils ne surviennent,
- utilise toute l'information disponible pour apporter des améliorations,
- n'est pas liée à une norme particulière en matière de qualité,
- s'applique principalement à l'étape de la planification,
- est une activité englobante

Contrôle de la qualité

- corrige les problèmes relevés,
- utilise des mesures permanentes pour prendre des décisions au sujet des processus ou des produits,
- exige une norme de qualité préétablie pour permettre la comparaison,
- s'applique principalement à l'étape du traitement,
- est une procédure sous-jacente à l'assurance de la qualité

3.7 Exercices

1. L'Académie Marguerite Bourgeois s'est vu accorder une subvention importante qui permettrait la construction d'une nouvelle bibliothèque ou d'un nouveau gymnase. Comme la subvention ne suffira que pour une seule installation, la directrice veut demander aux élèves laquelle, selon eux, a le plus besoin d'être remplacée.

Le tableau ci-dessous indique le nombre d'élèves de l'école selon le genre et l'année d'études, de l'éducation préscolaire à la 12e année (secondaire 5).

Tableau 3.7.1
Nombre d'élèves selon le sexe et l'année d'études, Académie Marguerite Bourgeois

	Préscolaire	1	2	3	4	5	6	7	8	9	10	11	12
Garçons	9	8	9	9	13	20	23	28	78	74	69	71	60
Filles	6	8	11	10	13	18	35	34	63	62	61	88	70
Total	15	16	20	19	26	38	58	62	141	136	130	159	130

Quelle est la population étudiante totale de l'Académie Marguerite Bourgeois?

c. La directrice veut échantillonner 50 % des élèves. De combien d'élèves serait constitué l'échantillon?

d. La directrice veut maintenir dans l'échantillon la bonne proportion de filles par rapport aux garçons. Calculez, à l'aide de la formule suivante, le nombre d'élèves de genre masculin de l'éducation préscolaire qu'il faudrait inclure dans l'échantillon.

$$\frac{nombre\ d'élèves\ de\ sexe\ masculin\ de\ la\ maternelle}{nombre\ total\ d'élèves} \times taille\ de\ l'échantillion$$

e. Quel type de technique d'échantillonnage a-t-on utilisé?

f. Si la directrice désire échantillonner 180 élèves, combien y aura-t-il de garçons et de filles par année d'études dans l'échantillon? Inscrivez vos réponses dans un tableau. (Arrondissez les résultats au nombre entier le plus près.)

2. D'après ce que vous savez déjà du processus statistique, placez les étapes suivantes dans le bon ordre :

- traitement
- collecte
- information
- données

3. Quelles sont les conséquences de l'absence de vérification des données sur l'information produite?

3.8 Réponses

1. a. La population étudiante totale de l'Académie Marguerite Bourgeois est de 950 élèves.
   b. Un échantillon de 50 % de la population étudiante de l'école équivaudrait à 475 élèves.
   c. Le nombre de garçons de l'éducation préscolaire qu'il faudrait inclure dans un échantillon de 475 est 4 ou 5. 9/950 x 475 = 4,5
   d. La méthode d'échantillonnage ici utilisée est l'échantillonnage stratifié.
   e. Le tableau qui suit présente la ventilation de l'échantillon de 180 élèves de façon à représenter proportionnellement les genres par année d'études.

Tableau 3.8.1
Nombre d'élèves qu'il faut pour constituer un échantillon de 180 élèves de l'Académie Marguerite Bourgeois, selon le genre et l'année d'étude

	Préscolaire	1	2	3	4	5	6	7	8	9	10	11	12
Garçons	2	2	2	2	2	4	4	5	15	14	13	13	11
Filles	1	2	2	2	2	3	7	6	12	12	12	17	13
Total	3	4	4	4	4	7	11	11	27	26	25	30	24

2. Le bon ordre de ces étapes est le suivant :

- données
- collecte
- traitement
- information

3. Les données non vérifiées peuvent contenir des erreurs et par conséquent être inexactes ou incomplètes. L'information inexacte fait l'objet d'une vérification avant sa diffusion au grand public.

4 Exploration des données

À plusieurs étapes du processus de production d'information statistique, il peut être utile d'explorer les données. Cela peut être au moment d'évaluer si une source de données répond à vos besoins, au moment où vous recevez les données brutes et voulez décider du traitement qui sera nécessaire pour pouvoir les utiliser ou avant de réaliser des analyses statistiques plus avancées. Peu importe la source des données, il est important de bien les comprendre et d'identifier les limites. Pour ce faire, vous pouvez vous poser les questions suivantes :

- Quelles sont les métadonnées disponibles pour cet ensemble de données? Les descriptions des variables sont-elles disponibles?
- Quelles sont la population observée, l'unité d'observation et la période de référence?
- S'agit-il de données agrégées ou de microdonnées?
- De quels types sont les variables présentes dans le fichier?
- Quelles sont les distributions de fréquences de ces variables? Quelles sont les mesures de tendance centrale et de dispersion?

Cette section commence par la présentation de quelques outils informatiques utiles pour explorer les données. Les différents types de variables sont ensuite présentés, suivis par les statistiques descriptives qui permettent d'explorer les données, c'est-à-dire les tableaux de fréquences et les mesures de tendance centrale et de dispersion.

4.1 Outils d'exploration des données

Les logiciels de production de graphiques, de programmation, de base de données et de tabulation sont régulièrement utilisés pour explorer les données. En voici quelques exemples :

- Les **tableurs** sont des programmes qui permettent d'additionner des colonnes et des lignes de nombres, de calculer des moyennes et de réaliser des analyses descriptives. On peut s'en servir pour produire des tableaux sommaires des résultats. Les tableurs permettent aussi de produire des graphiques pour mieux comprendre les relations entre les variables. Ceux-ci se présentent sous des formes diverses : graphiques à barres, graphiques linéaires, graphiques circulaires, pour ne nommer que quelques exemples de visualisations des données.
- Les données sont parfois sauvegardées dans des **bases de données** pour en faciliter l'accès et permettre la production de sommaires, de données agrégées et de rapports. Un logiciel de base de données devrait être en mesure d'enregistrer, de récupérer, de trier et d'analyser des données.
- Les **programmes spécialisés** peuvent servir à vérifier, à nettoyer, à imputer et à traiter le tableau final. Ils offrent tous les services en un seul module et peuvent servir après chaque cycle de la même enquête saisie dans le système. Ces programmes produisent par la suite les résultats prêts à être publiés.
- Les **logiciels statistiques** permettent à la fois de traiter les données, de produire des résultats sommaires et des visualisations, mais ils permettent en plus de réaliser des analyses statistiques avancées comme des modélisations.

Un exemple d'outil très populaire pour explorer les données est le logiciel R. Il s'agit d'un langage de programmation et d'un logiciel libre que tous peuvent télécharger et installer sur leur ordinateur pour manipuler, explorer et analyser les données. Les graphiques présentés dans les prochaines sections ont tous été créés à l'aide de R.

Les résultats obtenus à l'aide de ces différents outils peuvent servir de nombreuses façons. Ils peuvent être enregistrés en vue d'une récupération et d'une utilisation ultérieure, être transmis à d'autres équipes sous forme de fichiers électroniques ou être diffusés sur le web pour communiquer l'information statistique à ceux qui en ont besoin. Il s'agit généralement d'un auditoire précis et la transmission des résultats doit être pensée en fonction de ces utilisateurs. Il faut répondre aux questions suivantes :

- À qui sont destinés les résultats produits?
- Sous quel format les résultats seront-ils mieux compris?

4.2 Types de variables

Une variable est une caractéristique mesurable qui peut prendre différentes valeurs. La taille, l'âge, le revenu, la province ou le pays de naissance, les années d'études et le type de logement sont tous des exemples de variables. Les variables peuvent être classées en deux catégories principales : les catégoriques et les variables numériques. Chacune des catégories se sépare en deux sous-catégories : nominale et ordinales pour les variables catégoriques, discrètes et continues pour les variables numériques. Ces types sont définis brièvement dans cette section.

Variables catégoriques

Une variable catégorique (aussi appelée variable qualitative) réfère à une caractéristique qui n'est pas quantifiable. Une variable catégorique peut être nominale ou ordinale.

Variables nominales

Une variable nominale décrit un nom, une étiquette ou une catégorie sans ordre naturel. Le sexe et le genre de logement en sont des exemples. Dans le tableau 4.2.1, la variable « Mode de transport pour se rendre au travail » est également une variable nominale.

Tableau 4.2.1
Mode de transport habituel utilisé par les Canadiens pour se rendre au travail

Mode de transport pour se rendre au travail	Nombre de personnes
Automobile, camion ou fourgonnette (conducteur)	9 929 470
Automobile, camion ou fourgonnette (passager)	923 975
Transport en commun	1 406 585
À pied	881 085
Bicyclette	162 910
Autres moyens	146 835

Variables ordinales

Une variable ordinale est une variable dont les valeurs sont définies par une relation d'ordre entre les catégories possibles. Dans le tableau 4.2.2, la variable « comportement » est ordinale parce que la catégorie « Excellent » est meilleure que la catégorie « Très bon », qui est elle-même meilleure que la catégorie « Bon » et ainsi de suite. On y trouve un certain ordre naturel, mais celui-ci est limité par le fait que nous ne savons pas dans quelle mesure le comportement « Excellent » est meilleur que le comportement « Très bon » par exemple.

Tableau 4.2.2
Classement des élèves selon le comportement

Comportement	Nombre d'élèves
Excellent	5
Très bon	12
Bon	10
Mauvais	2
Très mauvais	1

Il est important de noter que bien que les variables catégoriques ne soient pas quantifiables, elles peuvent apparaître sous forme de nombre dans un ensemble de données. La correspondance entre ces nombres et les catégories correspondantes est établie au cours du codage des données. Pour bien identifier les types de variables, il faut donc s'assurer de disposer des métadonnées (les données à propos des données) qui doivent inclure les ensembles de codes utilisés pour chaque variable catégorique. Par exemple, les catégories présentées dans le tableau 4.2.2 pourraient apparaître sous forme d'un nombre allant de 1 à 5 : 1 pour « très mauvais », 2 pour « mauvais », 3 pour « bon », 4 pour « très bon » et 5 pour « excellent ».

Variables numériques

Une variable numérique (aussi appelée variable quantitative) est une caractéristique quantifiable dont les valeurs sont des nombres, à l'exclusion des nombres qui correspondent en fait à des codes. Les variables numériques peuvent être continues ou discrètes.

Variables continues

On dit qu'une variable est continue si elle prend un nombre infini de valeurs réelles possibles à l'intérieur d'un intervalle donné. Prenons la taille d'un élève par exemple. La taille ne peut pas prendre n'importe quelle valeur. Elle ne peut pas être négative, ni être plus grande que trois mètres. Mais le nombre de valeurs possibles que peut prendre la taille est théoriquement infini. Un élève pourrait mesurer 1,632 174 875 5… mètres par exemple. Il s'agit donc d'une variable continue. En pratique, les méthodes utilisées ou la précision des instruments employés pour mesurer une variable continue en restreignent la précision. La taille rapportée sera arrondie au centimètre près, soit 1,63 m. L'âge est un autre exemple de variable continue qui est le plus souvent rapportée en arrondissant à l'entier inférieur.

Variables discrètes

Contrairement à une variable continue, une variable discrète ne peut prendre qu'un nombre fini de valeurs réelles possibles à l'intérieur d'un intervalle donné. La note accordée par un juge à un gymnaste lors d'une compétition est un exemple de variable discrète : la plage varie de 0 à 10 et la note ne comporte jamais plus qu'une décimale (p. ex., une note de 8,5). On peut donc énumérer toutes les valeurs possibles (0, 0,1, 0,2…) et constater que le nombre de valeurs possibles est fini : il est de 101! Un autre exemple est la taille du ménage. Prenons les ménages qui ont 20 personnes ou moins. Le nombre de valeurs possibles dans cet intervalle sera de 20, car on sait qu'il n'est pas possible pour un ménage d'inclure un nombre de personnes qui serait une fraction d'un nombre entier comme 2,27 par exemple.

4.3 Distribution de fréquences

La **fréquence (f)** d'une valeur particulière est le nombre de fois que celle-ci se dégage des données. La **distribution** d'une variable est le profil des valeurs , c'est-à-dire l'ensemble formé de toutes les valeurs possibles et des fréquences associées à ces valeurs. Les distributions de fréquences sont représentées sous forme de tableaux ou de graphiques.

La **distribution de fréquences** peut indiquer soit le nombre réel d'observations s'inscrivant dans chaque intervalle ou le pourcentage d'observations. Dans le dernier cas, la distribution s'appelle une **distribution de fréquences relatives**.

Les tableaux de distribution de fréquences servent autant pour les variables catégoriques que pour les variables numériques. On ne devrait utiliser des variables continues qu'avec des intervalles de classe, comme nous l'expliquerons un peu plus loin.

Voyons quelques exemples de distribution de fréquences et de distribution de fréquences relatives à partir de variables discrètes.

Exemple 1 – Construction d'un tableau de distribution de fréquences

On a réalisé une enquête sur l'avenue des Érables. Dans chacune des 20 maisons, on a demandé aux gens d'indiquer le nombre de véhicules immatriculés dans leur ménage. Voici les résultats enregistrés :

1, 2, 1, 0, 3, 4, 0, 1, 1, 1, 2, 2, 3, 2, 3, 2, 1, 4, 0, 0

Suivez les étapes indiquées ci-dessous pour présenter ces données dans un tableau de distribution de fréquences.

1. Divisez les résultats (x) en intervalles, puis comptez le nombre de résultats dans chaque intervalle. Dans ce cas, les intervalles seraient le nombre de ménages n'ayant aucun véhicule (0), un véhicule (1), deux véhicules (2) et ainsi de suite.
2. Créez un tableau à l'aide de colonnes séparées pour les intervalles (le nombre de véhicules par ménage), les résultats cochés et la fréquence des résultats pour chaque intervalle. Intitulez ces colonnes Nombre de véhicules, Comptage et Fréquence.
3. Lisez la liste de données de gauche à droite et mettez une coche dans la rangée appropriée. Par exemple, le premier résultat est un 1; mettez donc une coche dans la rangée 1 de la colonne des intervalles (Nombre de véhicules). Le résultat suivant est un 2; mettez donc une coche dans la rangée 2 de la colonne des intervalles et ainsi de suite. Quand vous arriverez à votre cinquième coche, tracez la coche en travers des quatre coches précédentes pour faciliter la lecture de vos calculs finals de fréquence.
4. Additionnez le nombre de coches dans chaque rangée, puis enregistrez-les dans la dernière colonne, intitulée Fréquence.

Votre tableau de distribution de fréquences pour cet exercice devrait ressembler à ce qui suit :

Tableau de données du graphique 4.3.1

Borne supérieure de l'intervalle de classe du nombre de grimpeurs par jours	Fréquence relative cumulée (%)
9	3
19	10
29	20
39	37
49	57
59	87
69	100

0 zéro absolu ou valeur arrondie à zéro

Si nous examinons rapidement ce tableau de distribution de fréquences, nous pouvons constater que sur les 20 ménages sondés, 4 n'avaient pas de véhicule, 6 en avaient 1, etc.

Exemple 2 – Construction d'un tableau de distribution de fréquences cumulées

Un tableau de distribution de fréquences cumulées est un tableau plus détaillé. Il ressemble presque à un tableau de distribution de fréquences, mais on y ajoute des colonnes qui donnent la fréquence cumulée et le pourcentage cumulé des résultats.

Lors d'un récent tournoi d'échecs, les 10 participants ont rempli un formulaire sur lequel ils devaient indiquer leur nom, leur adresse et leur âge. Voici les âges des participants enregistrés :

36, 48, 54, 92, 57, 63, 66, 76, 66, 80

Suivez les étapes indiquées ci-dessous pour présenter ces données dans un tableau de distribution de fréquences cumulées.

1. Divisez les résultats en intervalles, puis comptez le nombre de résultats dans chaque intervalle. Dans ce cas, des intervalles de 10 sont convenables. Puisque 36 est le plus jeune âge et 92, le plus grand âge, commencez avec l'intervalle 35 à 44, puis terminez avec l'intervalle 85 à 94.

2. Créez un tableau semblable au tableau de distribution de fréquences, mais ajoutez trois colonnes additionnelles. Dans la première colonne, celle de la **Valeur inférieure**, indiquez la valeur inférieure des intervalles des résultats. Dans la première rangée, par exemple, vous indiqueriez le nombre 35.
    - La colonne suivante est la colonne **Valeur supérieure**. Indiquez la valeur supérieure des intervalles des résultats. Vous indiqueriez, par exemple, le nombre 44 dans la première rangée.
    - La troisième colonne est la colonne **Fréquence**. Enregistrez le nombre de fois qu'un résultat apparaît entre les valeurs inférieure et supérieure. Indiquez le chiffre 1 dans la première rangée.
    - La quatrième colonne est la colonne **Fréquence cumulée**. Il s'agit ici d'ajouter la fréquence indiquée dans la rangée précédente à la fréquence indiquée dans la rangée courante. Dans la première rangée, la fréquence cumulée est identique à la fréquence. Toutefois, dans la deuxième rangée, il faut ajouter la fréquence de l'intervalle 35 à 44 (1) à la fréquence de l'intervalle 45 à 54 (2). Ainsi, la fréquence cumulée est 3, c'est-à-dire, qu'on dénombre 3 participants dans le groupe d'âge des 34 à 54 ans.
      1 + 2 = 3
    - La colonne suivante est la colonne **Pourcentage**. Dans cette colonne, indiquez le pourcentage de la fréquence. Pour ce faire, divisez la fréquence par le nombre total de résultats, puis multipliez par 100. Dans ce cas, la fréquence de la première rangée est 1 et le nombre total de résultats est 10. Le pourcentage serait donc
      10,0. (1 ÷ 10) X 100 = 10,0
    - La dernière colonne est la colonne **Pourcentage cumulé**. Dans cette colonne, divisez la fréquence cumulée par le nombre total de résultats, puis multipliez par 100. Notez que le dernier nombre dans cette colonne devrait toujours être égal à 100,0. Dans cet exemple, la fréquence cumulée est 1 et le nombre total de résultats est 10; le pourcentage cumulé de la première rangée est donc
      10,0. (1 ÷ 10) X 100 = 10,0

Le tableau de distribution de fréquences cumulées devrait ressembler à ce qui suit :

Tableau 4.3.2
Âges des participants à un tournoi d'échecs

Valeur inférieure	Valeur supérieure	Fréquence (f)	Fréquence cumulée	Pourcentage	Pourcentage cumulé
35	44	1	1	10	10
45	54	2	3	20	30
55	64	2	5	20	50
65	74	2	7	20	70
75	84	2	9	20	90
85	94	1	10	10	100

Intervalles de classe

Si une variable revêt un plus grand nombre de valeurs, il est alors plus facile de présenter et de manipuler les données en groupant les valeurs dans des intervalles de classe. On présente toujours les variables continues en intervalles de classe, tandis qu'on peut choisir de grouper ou de ne pas grouper les valeurs discrètes dans des intervalles de classe.

Pour illustrer notre propos, supposons que nous définissons des groupes d'âge pour une étude portant sur les jeunes, en tenant compte de la possibilité d'inclure certaines personnes plus âgées.

La fréquence d'un intervalle de classe est le nombre d'observations comprises dans un intervalle prédéfini particulier. Par exemple, si les données de notre étude indiquent que 20 personnes sont âgées de 5 à 9 ans, la fréquence de l'intervalle 5 à 9 sera 20.

Les extrémités d'un intervalle de classe sont les valeurs les plus faibles et les plus élevées qu'une variable peut revêtir. Ainsi, les intervalles dans notre étude sont 0 à 4 ans, 5 à 9 ans, 10 à 14 ans, 15 à 19 ans, 20 à 24 ans et 25 ans et plus. Les extrémités du premier intervalle sont 0 et 4, si la variable est discrète, et 0 et 4,999, si la variable est continue. Les extrémités des autres intervalles de classe seraient déterminées de la même façon.

La longueur d'un intervalle de classe est la différence entre l'extrémité inférieure d'un intervalle et l'extrémité inférieure de l'intervalle suivant. Ainsi, si les intervalles de notre étude sont 0 à 4, 5 à 9, etc., la longueur des cinq premiers intervalles est 5 et le dernier intervalle est ouvert. Les intervalles pourraient aussi être écrits sous la forme 0 à moins de 5, 5 à moins de 10, 10 à moins de 15, 15 à moins de 20, 20 à moins de 25 et 25 et plus.

Règles relatives aux ensembles de données qui renferment un grand nombre d'observations

En résumé, suivez ces règles de base lorsque vous construisez un tableau de distribution de fréquences pour un ensemble de données qui renferme un grand nombre d'observations :

- trouvez la valeur la plus faible et la valeur la plus élevée des variables,
- décidez de la longueur des intervalles de classe,
- incluez toutes les valeurs possibles de la variable.

Lorsqu'on décide de la longueur des intervalles de classe, il faut trouver un compromis afin d'avoir des intervalles assez courts (pour éviter que toutes les observations ne tombent dans le même intervalle) mais assez longs aussi (pour ne pas se retrouver avec une seule observation par intervalle).

Il importe aussi de veiller à ce que les intervalles de classe soient mutuellement exclusifs.

Exemple 3 – Construction d'un tableau de distribution de fréquences pour un grand nombre d'observations

On a testé 30 piles AA pour déterminer combien de temps elles dureraient. Voici les résultats du test, arrondis à la minute :

423, 369, 387, 411, 393, 394, 371, 377, 389, 409, 392, 408, 431, 401, 363, 391, 405, 382, 400, 381, 399, 415, 428, 422, 396, 372, 410, 419, 386, 390

Suivez les étapes indiquées dans l'exemple 1 et les règles ci-dessus pour vous aider à construire un tableau de distribution de fréquences.

Réponse

La valeur la plus faible est 363 et la valeur la plus élevée est 431.

Si l'on utilise les données fournies et un intervalle de classe de 10, l'intervalle de la première classe est 360 à 369 et inclut 363 (la valeur la plus faible). Souvenez-vous qu'il devrait toujours y avoir assez d'intervalles de classe pour que la valeur la plus élevée y soit incluse.

Le tableau de distribution de fréquences, une fois rempli, ressemble à ce qui suit :

Tableau 4.3.3
Durée de vie des piles AA, en minutes

Durée de vie des piles, en minutes (x)	Fréquence (f)
360 à 369	2
370 à 379	3
380 à 389	5
390 à 399	7
400 à 409	5
410 à 419	4
420 à 429	3
430 à 439	1
Total	**30**

Exemple 4 – Construction d'un tableau de fréquences relative et de fréquence en pourcentage

Un analyste qui étudierait les données de l'exemple 3 voudrait peut-être savoir non seulement combien de temps durent les piles, mais également quelle proportion d'entre elles s'inscrit à l'intérieur de chaque intervalle de classe de leur durée de vie.

On trouve la **fréquence relative** d'une observation particulière ou d'un intervalle de classe particulier en divisant la fréquence (f) par le nombre d'observations (n), c'est-à-dire (f ÷ n). Ainsi :

Fréquence relative = fréquence ÷ nombre d'observations

On trouve la **fréquence en pourcentage** en multipliant la valeur de chaque fréquence relative par 100. Ainsi :

Fréquence en pourcentage = fréquence relative X 100 = f ÷ n X 100

Utilisez les données fournies dans l'exemple 3 pour créer un tableau qui donnera la fréquence relative et la fréquence en pourcentage de chaque intervalle de classe de la vie des piles.

Voici ce à quoi ressemble ce tableau :

Tableau 4.3.4
Durée de vie des piles AA, en minutes

Durée de vie des piles, en minutes (x)	Fréquence (f)	Fréquence relative	Fréquence en pourcentage
360 à 369	2	0,07	7
370 à 379	3	0,1	10
380 à 389	5	0,17	17
390 à 399	7	0,23	23
400 à 409	5	0,17	17
410 à 419	4	0,13	13
420 à 429	3	0,1	10
430 à 439	1	0,03	3
Total	**30**	**1**	**100**

Un analyste qui examinerait ces données pourrait maintenant dire que :

- 7 % des piles AA ont une durée de vie d'au-moins 360 minutes, mais de moins de 370 minutes;
- la probabilité qu'une pile AA sélectionnée au hasard ait une durée de vie s'inscrivant à l'intérieur de cette plage est d'environ 0,07.

Exemple 5 – Visualisation de la distribution de fréquence relative cumulée

Comme nous l'avons vu à l'exemple 2, la distribution de fréquence cumulée est utilisée pour déterminer le nombre d'observations qui se situent au-dessous d'une valeur particulière dans un ensemble de données. Elle est calculée sur chaque ligne d'un tableau de fréquence en ajoutant à chaque fréquence la somme des fréquences sur les lignes qui précèdent. La dernière valeur sera toujours égale au total des observations, puisque toutes les fréquences auront déjà été ajoutées au total précédent. Voyons un exemple supplémentaire de calcul de la fréquence cumulée.

On a compté et enregistré durant une période de 30 jours le nombre de gens qui faisaient de l'escalade autour du lac Louise, en Alberta. Voici les résultats du décompte :

31, 49, 19, 62, 24, 45, 23, 51, 55, 60, 40, 35 54, 26, 57, 37, 43, 65, 18, 41, 50, 56, 4, 54, 39, 52, 35, 51, 63, 42.

Le nombre de grimpeurs varie de 4 à 65. Pour créer le tableau de fréquences, il vaut mieux grouper les données en intervalles de classe de 10. Chaque intervalle peut être une ligne dans le tableau de fréquence. La colonne **Fréquence** sert à indiquer le nombre d'observations qui se situent à l'intérieur d'un intervalle de classe. Par exemple, Il n'y a que deux valeurs dans l'intervalle de 10 à 20, alors sa fréquence est de 2 dans le tableau correspondant.

Utilisez la colonne **Fréquence** pour le calcul de la fréquence cumulée.

1. Premièrement, ajoutez le nombre tiré de la colonne **Fréquence** au nombre précédent. Dans la première ligne, par exemple, nous n'avons qu'une seule observation et aucun nombre qui précède. La fréquence cumulée est donc un.
   1 + 0 = 1
2. Dans la seconde ligne, cependant, il y a deux observations. Ajoutez ces deux observations à la fréquence cumulée précédente (un) et vous obtiendrez trois.
   1 + 2 = 3
3. Enregistrez les résultats dans la colonne **Fréquence cumulée**.

Les autres entrées du tableau peuvent être calculées de manière similaire. Les résultats obtenus sont présentés au tableau 4.3.5.

Tableau 4.3.5
Fréquence et fréquence cumulée du nombre de grimpeurs par jour autour du lac Louise, en Alberta, pendant une période de 30 jours

Nombre de grimpeurs	Fréquence (f)	Fréquence cumulée
<10	1	1
10 à <20	2	1 + 2 = 3
20 à <30	3	3 + 3 = 6
30 à <40	5	6 + 5 = 11
40 à <50	6	11 + 6 = 17
50 à <60	9	17 + 9 = 26
>= 60	4	26 + 4 = 30

La distribution de fréquence relative cumulée est une autre façon de représenter une distribution de fréquences. Elle consiste à calculer le pourcentage de la fréquence cumulée dans chaque intervalle.

On calcule la fréquence relative cumulée en divisant la fréquence cumulée par le nombre total d'observations (n), qu'on multiplie ensuite par 100 (la dernière valeur est toujours égale à 100 %). Ainsi :

Fréquence relative cumulée = (fréquence cumulée ÷ n) x 100

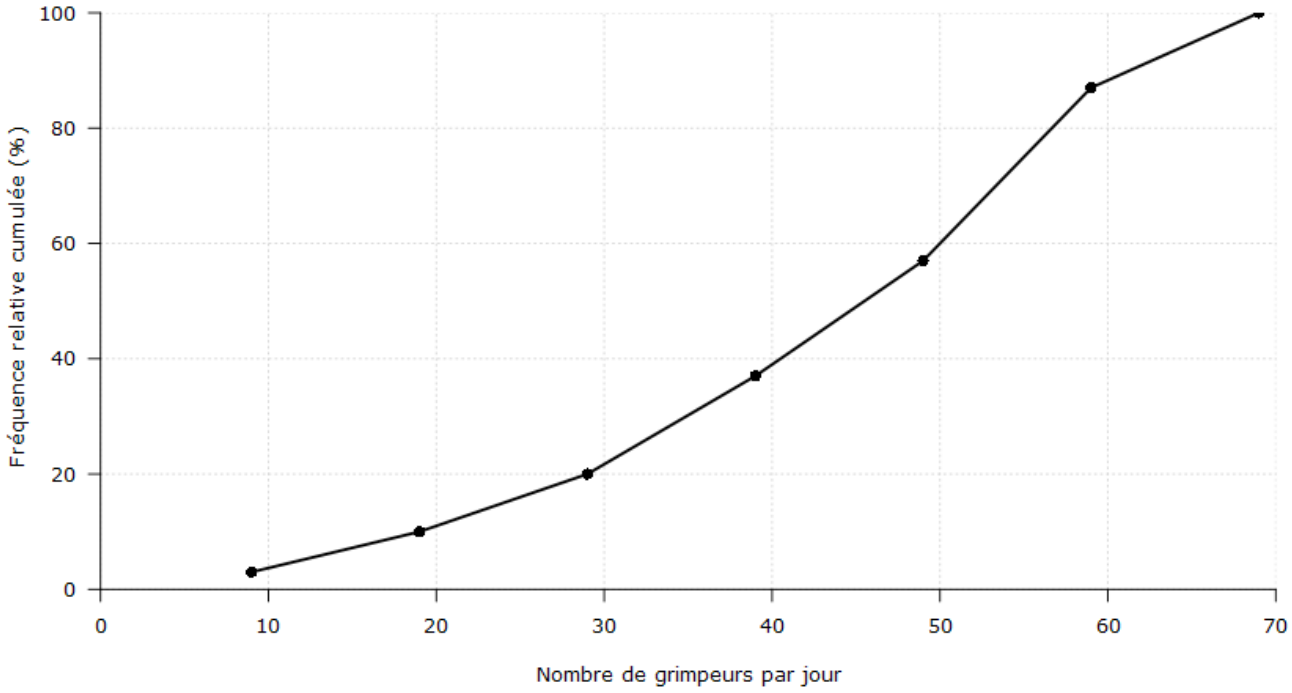
La quatrième colonne du tableau 4.3.6 illustre le calcul de la fréquence relative cumulée du nombre de grimpeurs par jour au lac Louise.

Tableau 4.3.6
Fréquence relative cumulée du nombre de grimpeurs par jour autour du lac Louise, en Alberta, pendant une période de 30 jours

Nombre de grimpeurs	Fréquence (f)	Fréquence cumulée	Fréquence relative cumulée (%)
<10	1	1	1 ÷ 30 x 100 = 3
10 à <20	2	1 + 2 = 3	3 ÷ 30 x 100 = 10
20 à <30	3	3 + 3 = 6	6 ÷ 30 x 100 = 20
30 à <40	5	6 + 5 = 11	11 ÷ 30 x 100 = 37
40 à <50	6	11 + 6 = 17	17 ÷ 30 x 100 = 57
50 à <60	9	17 + 9 = 26	26 ÷ 30 x 100 = 87
>= 60	4	26 + 4 = 30	30 ÷ 30 x 100 = 100

La distribution de fréquence relative cumulée peut être visualisée à l'aide d'un graphique à barres ou d'un graphique linéaire, comme au graphique 4.3.1 ci-dessous. La valeur sur l'axe horizontal correspond à la borne supérieure de l'intervalle de classe.

Graphique 4.3.1
Fréquence relative cumulée du nombre de grimpeurs par jour autour du lac Louise, Alberta, au cours d'une période de 30 jours



On peut voir au graphique 4.3.1 qu'au cours de la majorité (57 %) des jours de la période, le nombre de grimpeurs par jour a été inférieur ou égal à 49.

Une distribution de fréquence peut être visualisée à l'aide :

- d'un graphique circulaire (variable nominale),
- d'un graphique à barres (variable nominale ou ordinale),
- d'un graphique linéaire (variable ordinale ou discrète),
- ou d'un histogramme (variable continue).

Ces types de graphiques seront présentés plus en détail dans la section 5 consacrée à la visualisation des données. Mais d'abord nous verrons d'autres méthodes pour résumer les données à l'aide des mesures de tendance centrale et de dispersion.

4.4 Mesures de la tendance centrale

La meilleure façon de résumer un ensemble de données par une seule valeur est de trouver la valeur la plus représentative, celle qui indique où se situe le centre de la distribution. C'est ce que l'on appelle la tendance centrale. Les trois mesures de tendance centrale les plus courantes sont :

- La **moyenne arithmétique**, qui consiste à calculer la somme des valeurs et la diviser par le nombre de valeurs,
- La **médiane**, qui est le point milieu lorsque toutes les valeurs sont classées par ordre croissant,
- Le **mode**, qui est la valeur la plus typique de l'ensemble, c'est-à-dire celle qui apparaît le plus souvent.

Dans les prochaines sections, la façon de calculer ces trois mesures sera expliquée à l'aide d'exemples.

4.4.1 Calcul de la moyenne

La moyenne est calculable pour les variables numériques, qu'elles soient discrètes ou continues. On l'obtient simplement en additionnant l'ensemble des valeurs et en divisant cette somme par le nombre de valeurs. Ce calcul peut être fait à partir des données brutes ou d'un tableau de fréquences. Voici quelques exemples de calcul.

Exemple 1 – Tournoi de soccer au Mont Rival

Le Mont Rival organise un tournoi de soccer une fois par année. Au cours de la présente saison, le marqueur en tête de l'équipe hôte a compté 7, 5, 0, 7, 8, 5, 5, 4, 1 et 5 buts en dix parties. Quelle était sa moyenne de buts comptés?

La somme de ces valeurs est égale à 47 et il y a 10 valeurs. La moyenne est donc de 47 ÷ 10 = 4,7 buts par partie.

Exemple 2 – Accidents mortels de la route

Le tableau qui suit indique le nombre de personnes décédées dans des accidents de la route au cours d'une période de 10 ans. Durant cette période, quel a été le nombre moyen de personnes ayant perdu la vie annuellement sur les routes? Combien de personnes sont mortes en moyenne chaque jour dans des accidents de la route durant la même période?

Tableau 4.4.1.1
Nombre de décès dus aux accidents de la route

Année	Décès
2009	623
2010	583
2011	959
2012	1 037
2013	960
2014	797
2015	663
2016	652
2017	560
2018	619
Total	**7 453**

Le nombre total de décès est présenté dans le tableau (7 453). Il suffit de le diviser par 10 car le tableau couvre 10 années et on obtient une moyenne de 745,3 décès par année. Pour obtenir la moyenne quotidienne, il suffit de diviser par 365, ce qui donne environ 2 décès par jour.

Pour un jeu de données plus grand, il peut être plus pratique de commencer par résumer les données dans un tableau de fréquences avant de calculer la moyenne. Il faudra alors utiliser une somme pondérée par la fréquence de chaque valeur, comme dans l'exemple 3.

Exemple 3 – Tournoi de soccer au Mont Rival (Partie 2)

Retournons au tournoi de soccer du Mont Rival. Supposons que cinq équipes s'affrontaient dans ce tournoi, chacune d'elle incluant 10 joueurs pour un total de 50 joueurs. Le nombre de buts comptés par chaque joueur a été compilé, puis résumé dans le tableau de fréquence ci-dessous. Par exemple, on peut voir que 8 joueurs ont compté un seul but au cours du tournoi. Quel est le nombre moyen de buts comptés par joueur au cours de ce tournoi?

Tableau 4.4.1.2
Nombre de joueurs par nombre de buts comptés

Nombre de buts comptés	Nombre de joueurs
0	2
1	8
2	14
3	12
4	8
5	4
6	2

0 zéro absolu ou valeur arrondie à zéro

Il faut d'abord calculer le nombre total de buts comptés. Pour cela, il faut prendre chaque valeur observée du nombre de buts comptés, soit les valeurs 0 à 6, et les multiplier par le nombre de joueurs ayant réalisé ce nombre de buts :

$0 \times 2 + 1 \times 8 + 2 \times 14 + 3 \times 12 + 4 \times 8 + 5 \times 4 + 6 \times 2 = 136$

Puisqu'il y a 50 joueurs, la moyenne est donc de $136 \div 50 = 2{,}72$ buts par joueur.

4.4.2 Calcul de la médiane

La médiane est le point milieu d'un jeu de données, de sorte que 50 % des unités ont une valeur inférieure ou égale à la médiane et 50 % des unités ont une valeur supérieure ou égale. Dans un jeu de données de petite taille, il suffit de compter le nombre de valeurs (n) et de les ordonner en ordre croissant. Si le nombre de valeurs est un nombre impair, il faut lui additionner 1, puis le diviser par 2 pour obtenir le rang qui correspondra à la médiane. Le rang est la position d'une valeur une fois l'ensemble ordonné : la plus petite valeur correspond au rang 1, la seconde plus petite valeur au rang 2, etc.

Exemple 1 – Temps médian au 200 mètres d'un champion de course

Supposons qu'un champion de course effectue une course d'entraînement typique de 200 mètres dans les temps suivants : 26,1 secondes, 25,6 secondes, 25,7 secondes, 25,2 secondes, 25,0 secondes, 27,8 secondes et 24,1 secondes. Comment calcule-t-on le temps médian?

Commençons par classer les valeurs en ordre croissant.

Tableau 4.4.2.1
Rang associé à chaque valeur du temps au 200 mètres

Rang	Temps (en secondes)
1	24,1
2	25,0
3	25,2
4	25,6
5	25,7
6	26,1
7	27,8

Il y a n = 7 valeurs, un nombre impair. La médiane correspondra donc à la valeur de rang

$(n+1) \div 2 = (7 +1) \div 2 = 4$

Le temps médian est de 25,6 secondes.

Si le nombre de valeurs est un nombre pair, la médiane correspondra à la moyenne des valeurs de rang $n \div 2$ et $(n \div 2) +1$.

Exemple 2 – Temps médian au 200 mètres d'un champion de course (Partie 2)

Maintenant, supposons que le coureur effectue sa huitième course de 200 mètres en 24,7 secondes. Dans ce cas, quelle est la valeur médiane?

Tableau 4.4.2.2
Rang associé à chaque valeur du temps au 200 mètres, mise à jour

Rang	Temps (en secondes)
1	24,1
2	24,7
3	25,0
4	25,2
5	25,6
6	25,7
7	26,1
8	27,8

Il y a maintenant n = 8 valeurs, un nombre pair. La médiane correspondra à la moyenne entre la valeur de rang

$n \div 2 = 8 \div 2 = 4$

et la valeur de rang

$(n \div 2) + 1 = (8 \div 2) +1 = 5$

Le temps médian est donc de $(25,2 +25,6) \div 2 = 25,4$ secondes.

Pour les ensembles de données plus grands, il est possible d'utiliser la distribution de fréquence relative cumulée pour aider à identifier la médiane. La médiane sera la plus petite valeur pour laquelle la fréquence relative cumulée atteint au moins 50 %. Il est toutefois mieux d'utiliser une fonction statistique de base disponible dans un tableur ou un logiciel statistique, car le résultat sera plus fiable. Voyons un exemple.

Exemple 3 – Taille médiane du ménage des élèves de la classe

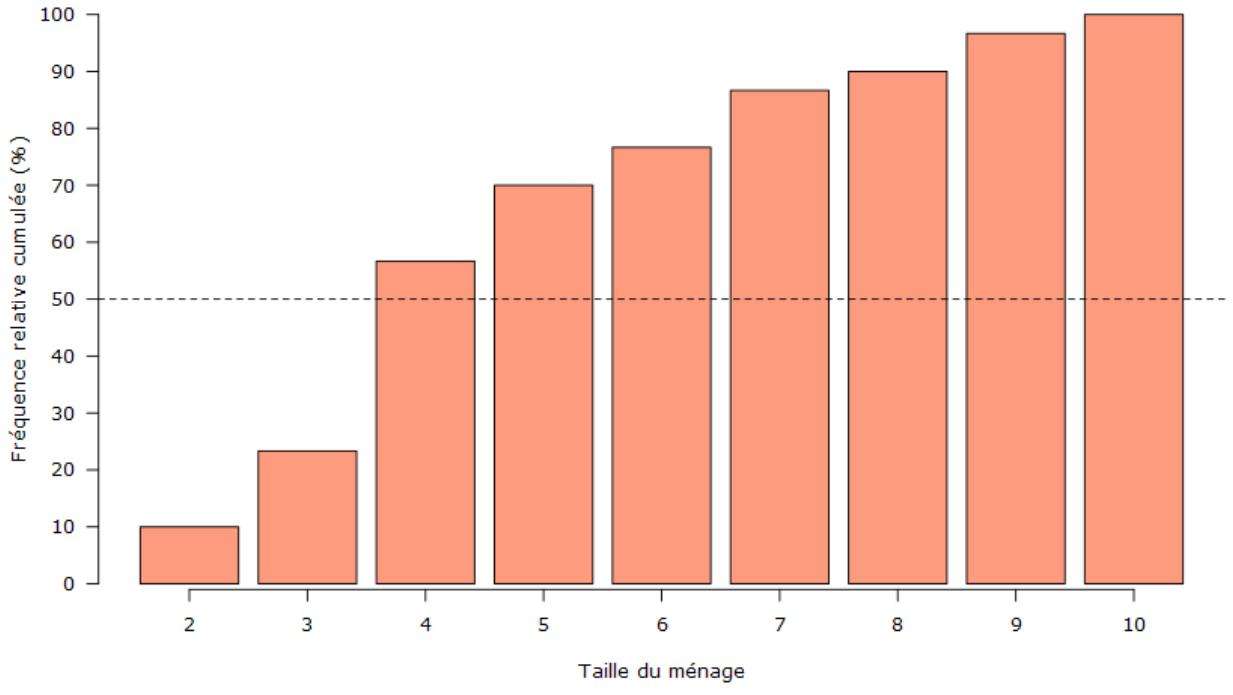
Supposons que vous demandez aux 30 élèves de votre classe combien de personnes vivent dans leur ménage. Vous résumez les données recueillies dans un tableau de fréquences, dans lequel vous incluez la fréquence relative et la fréquence relative cumulée.

Tableau 4.4.2.3
Tableau de fréquence de la taille du ménage des élèves

Taille du ménage	Fréquence (nombre d'élèves)	Fréquence relative (%)	Fréquence cumulée (nombre d'élèves)	Fréquence relative cumulée (%)
2	3	10,0	3	10,0
3	4	13,3	7	23,3
4	10	33,3	17	56,7
5	4	13,3	21	70,0
6	2	6,7	23	76,7
7	3	10,0	26	86,7
8	1	3,3	27	90,0
9	2	6,7	29	96,7
10	1	3,3	30	100,0

Vous pouvez voir que 10,0 % des élèves (3 élèves) vivent dans un ménage de taille 2, 23,3 % des élèves (7 élèves) vivent dans un ménage de taille 3 ou moins et 56,7 % des élèves (17 élèves) vivent dans un ménage de taille 4 ou moins. La médiane sera égale à 4, car c'est la plus petite valeur pour laquelle la fréquence cumulée dépasse 50 %. Ceci est encore plus évident si on visualise la fréquence relative cumulée grâce à un graphique à barres comme le graphique 4.4.2.1 ci-dessous. La ligne pointillée y indique la fréquence relative cumulée de 50 %.

Graphique 4.4.2.1
Fréquence relative cumulée de la taille des ménages des élèves de la classe



La moyenne, quant à elle, est obtenue en divisant le nombre total de personne dans les ménages des élèves de la classe :

$2 \times 3 + 3 \times 4 + 4 \times 10 + 5 \times 4 + 6 \times 2 + 7 \times 3 + 8 \times 1 + 9 \times 2 + 10 \times 1 = 147$

par le nombre d'élèves (30). Elle est donc de $147 \div 30 = 4{,}9$ personnes par ménage.

Dans cet exemple, la médiane (4) est un peu plus petite que la moyenne (4,9).

L'avantage d'utiliser la médiane plutôt que la moyenne est qu'elle est plus robuste aux valeurs extrêmes qui pourraient surgir à l'une des extrémités de la distribution. Il est donc important de vérifier si les données comptent des valeurs extrêmes avant de choisir quelle mesure de tendance centrale doit être utilisée. Ceci sera illustré par l'exemple ci-dessous.

Exemple 4 – Taille médiane du ménage des élèves de la classe (Partie 2)

Un nouvel élève est récemment inscrit dans votre classe. Vous décidez de vous renseigner auprès de lui sur la taille de son ménage afin de mettre vos résultats à jour. Il vous répond qu'il habite dans une très grande maison multigénérationnelle, qui compte 18 résidents!

La valeur de la moyenne après la mise à jour sera de $(147 + 18) \div 31 = 5{,}3$ personnes par ménage. Cet élève à lui seul a fait augmenter la moyenne de 0,4 personnes par ménage (5,3 - 4,9 = 0,4). Quant à la médiane, elle est restée inchangée. En effet, la fréquence relative cumulée pour la valeur 3 est de $7 \div 31 = 22{,}6$ % et celle pour la valeur 4 de $17 \div 31 = 54{,}8$ %. La valeur 4 est encore la plus petite valeur à atteindre une fréquence relative cumulée d'au moins 50 %.

4.4.3 Calcul du mode

Lorsqu'il est unique, le mode est la valeur d'une variable la plus souvent observée dans un ensemble de données et il peut alors être considéré comme une mesure de tendance centrale, au même titre que la moyenne et la médiane. Il est toutefois possible qu'il n'y ait aucun mode ou qu'il y ait plusieurs modes.

Il n'y a aucun mode lorsque toutes les valeurs possibles apparaissent le même nombre de fois dans l'ensemble de données. Il y a plusieurs modes lorsque la fréquence la plus élevée a été observée pour plusieurs valeurs différentes. Dans les cas où il n'y a aucun mode ou plusieurs modes, le mode ne peut pas être utilisé pour situer le centre de la distribution.

Le mode peut être utilisé pour résumer des variables catégoriques, alors que la moyenne et la médiane ne peuvent être calculées que pour les variables numériques. C'est d'ailleurs le principal avantage de cette mesure. Il est aussi utile pour les variables discrètes et pour les variables continues lorsqu'elles sont présentées par intervalles.

Voici quelques exemples de calcul du mode pour une variable discrète.

Exemple 1 – Nombre de buts dans un tournoi de hockey

Lors d'un tournoi de hockey, Audrey a compté 7, 5, 0, 7, 8, 5, 5, 4, 1 et 5 buts en dix parties. Une fois les données résumées dans un tableau de fréquences, il est facile de voir que le mode du nombre de buts d'Audrey sur les 10 parties est de 5, puisque cette valeur apparaît le plus souvent (4 fois). Le mode peut être considéré comme une mesure de tendance centrale pour cet ensemble de données, car il est unique.

Tableau 4.4.3.1
Nombre de parties selon le nombre de buts comptés

Nombre de buts	Fréquence (nombre de parties)
0	1
1	1
4	1
5	4
7	2
8	1

0 zéro absolu ou valeur arrondie à zéro

Exemple 2 – Nombre de points en douze parties de basket-ball

En 12 parties de basket-ball, Gabriel a compté 14, 14, 15, 16, 14, 16, 16, 18, 14, 16, 16 et 14 points. Une fois les données résumées dans un tableau de fréquences, il est facile de voir que cet ensemble de données est bimodal, c'est-à-dire qu'il y a deux modes, 14 et 16, parce que ce sont les valeurs les plus souvent observées (elles apparaissent 5 fois chacune). Le mode ne peut pas être utilisé comme mesure de tendance centrale, car il y en a plus d'un. Il s'agit d'une distribution bimodale.

Tableau 4.4.3.2
Nombre de parties selon le nombre de points comptés

Nombre de points	Fréquence (nombre de parties)
14	5
15	1
16	5
18	1

Exemple 3 – Nombre de touchés au cours de la saison de football

L'ensemble de données qui suit représente le nombre de touchés que Jérôme a marqué au cours de la saison de football : 0, 0, 1, 0, 0, 2, 3, 1, 0, 1, 2, 3, 1, 0. Comparons la moyenne, la médiane et le mode.

La somme ces valeurs est de 14 et le nombre de valeurs est de 14. La moyenne est donc égale à 1. Étant donné que le nombre de données est pair, la médiane sera égale à la moyenne entre la 7e et la 8e valeur une fois les données ordonnées :

Tableau 4.4.3.3
Rang associé aux valeurs du nombre de touchés à chaque partie

Rang	Nombre de touchés
1	0
2	0
3	0
4	0
5	0
6	1
7	1
8	1
9	1
10	1
11	2
12	2
13	3
14	3

Elle est donc égale à 1. Une fois les données résumées dans un tableau de fréquences, on peut voir que le mode est quant à lui égal à 0.

Tableau 4.4.3.4
Nombre de parties selon le nombre de touchés marqués

Nombre de touchés	Fréquence (nombre de parties)
0	6
1	4
2	2
3	2

0 zéro absolu ou valeur arrondie à zéro

En résumé, dans cet exemple, la moyenne est égale à 1, la médiane est égal à 1 et le mode est égal à 0.

Le mode est moins utilisé pour les variables continues, car il est probable qu'aucune valeur ne revienne plus d'une fois. Par exemple, si on demande à 20 personnes leur revenu exact au dollar près au cours de la dernière année, il se peut que certains aient des revenus très près les uns des autres, mais qu'ils ne soient jamais exactement le même. On peut alors utiliser des intervalles et déterminer la classe modale. L'histogramme permet de visualiser facilement la classe modale d'une variable continue.

Exemple 4 – Taille des personnes présentes à une partie de basketball

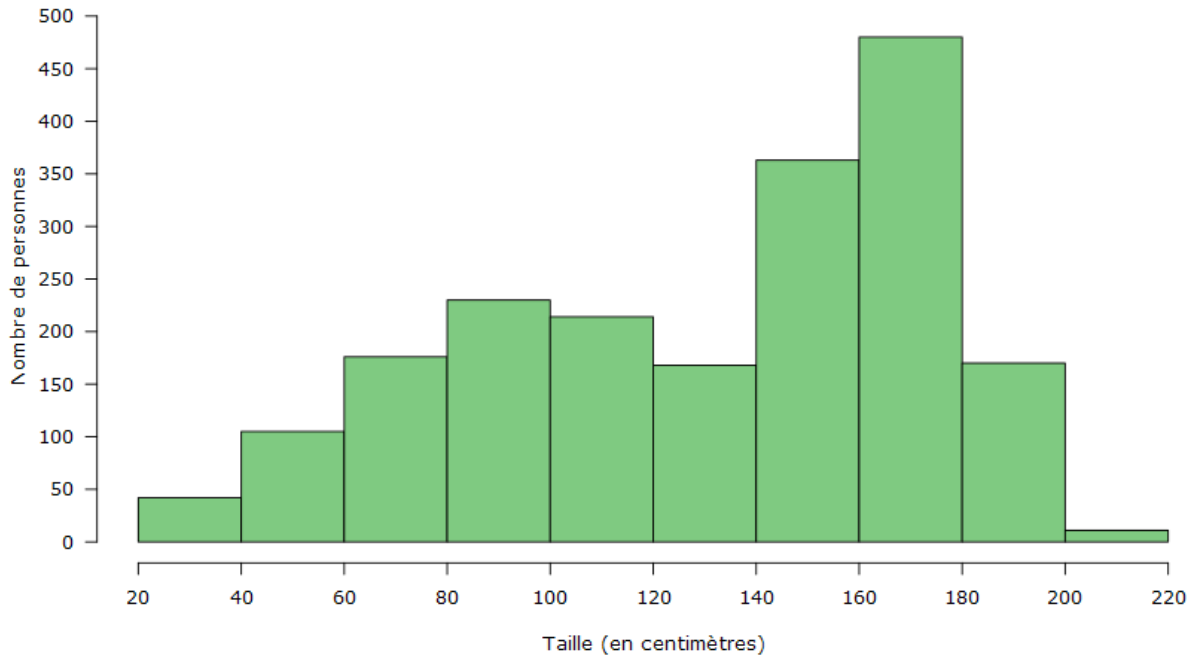
On s'intéresse à la taille des personnes présentes dans l'aréna lors d'une partie de basketball. Le tableau 4.4.3.5 présente le nombre de personnes pour chaque intervalle de 20 centimètres de la taille dans un jeu de données. La classe modale serait l'intervalle de 160 à 179 cm, car c'est celle avec la fréquence la plus élevée.

Tableau 4.4.3.5
Nombre de personnes par intervalle de taille

Taille (en centimètre)	Fréquence (nombre de personnes)
20 à 39	42
40 à 59	105
60 à 79	176
80 à 99	230
100 à 119	214
120 à 139	168
140 à 159	363
160 à 179	480
180 à 200	170
200 à 219	11

Le graphique 4.4.3.1 présente ces données sous la forme d'un histogramme.

Graphique 4.4.3.1
Histogramme de la taille des personnes présentes à la partie de basketball



Pour les variables catégoriques ou discrètes, les modes multiples correspondent à la même fréquence (la plus grande). Pour les variables continues, tous les sommets de la distribution peuvent être considérés comme des modes, même lorsqu'ils correspondent à des hauteurs différentes sur l'histogramme. La distribution de cet exemple serait donc bimodale, avec un mode majeur au niveau de la classe 160 à 179 et un mode mineur situé au niveau de la classe 80 à 99. La classe modale ne doit pas être utilisée comme mesure de tendance centrale, mais la découverte de deux modes nous donne une indication qu'il existe possiblement deux groupes distincts dans les données qui devraient être analysés séparément.

4.5 Mesures de la dispersion

Les mesures de tendance centrale visent à identifier la valeur la plus représentative d'un ensemble de données, c'est-à-dire le centre de la distribution. Pour obtenir une meilleure description d'un ensemble de données, il faut également une mesure de l'étalement des valeurs autour du centre. C'est ce qu'on appelle la dispersion. Les mesures de dispersion principales sont les suivantes :

- L'**étendue**, qui est la différence entre la plus petite valeur et la plus grande;
- L'**écart interquartile**, qui est l'étendue du 50% des données qui sont au centre de la distribution;
- La **variance**, qui est l'écart élevé au carré moyen entre chaque donnée et le centre de la distribution ;
- L'**écart-type**, la racine carrée de la variance.

Les sections qui suivent expliquent comment calculer ces mesures à l'aide d'exemples. Les mesures de dispersions s'appliquent uniquement aux variables numériques.

4.5.1 Calculer l'étendue et l'écart interquartile

Pour calculer l'étendue, il suffit de trouver la plus grande valeur observée d'une variable (le maximum) et de lui soustraire la plus petite valeur observée (le minimum). L'étendue ne tient compte que de ces deux valeurs et ignore les points de données entre les deux extrémités de la distribution. Elle sert de supplément à d'autres mesures, mais elle est rarement utilisée comme seule mesure de dispersion étant donné qu'elle est sensible aux valeurs extrêmes.

L'écart interquartile et l'écart semi-interquartile donnent une idée plus juste de la dispersion des données. Pour calculer ces deux mesures, il faut d'abord identifier les quartiles. Le quartile inférieur, ou premier quartile (Q1), est la valeur au-dessous de laquelle se trouvent 25 % des données lorsqu'elles sont arrangées en ordre croissant. Le quartile supérieur, ou troisième quartile (Q3), est la valeur au-dessous de laquelle se trouvent 75 % des données arrangées en ordre croissant. La médiane est considérée comme le second quartile (Q2). L'écart interquartile est la différence entre le quartile supérieur et le quartile inférieur. L'écart semi-interquartile est la moitié de l'écart interquartile.

Lorsque le jeu de données est petit, il est simple de trouver les valeurs des quartiles. Regardons un exemple.

Exemple 1 – Étendue et écart interquartile d'un ensemble de données

Identifiez les quartiles de l'ensemble de données suivant : 6, 47, 49, 15, 43, 41, 7, 39, 43, 41, 36.

Pour commencer, vous devez arranger les valeurs en ordre croissant. Ce faisant, vous pouvez donner un rang aux points de données. Le point correspondant à la plus petite valeur aura le rang 1, le point correspondant à la seconde plus petite valeur aura le rang 2 et ainsi de suite.

Tableau 4.5.1.1
Rang des points de données

Rang	Valeur
1	6
2	7
3	15
4	36
5	39
6	41
7	41
8	43
9	43
10	47
11	49

Il vous faut ensuite trouver le rang de la médiane. Comme vu à la section sur la médiane, lorsque le nombre de points est impair, la médiane correspond à la valeur du point de rang

(n + 1) ÷ 2 = (11 + 1) ÷ 2 = 6

La médiane est le point de données de rang 6. Il y a donc 5 valeurs de chaque côté.

Vous devez séparer la moitié inférieure à la médiane en 2. Le quartile inférieur sera donc la valeur du point de rang (5 +1) ÷2 = 3, ce qui donne Q1=15. La moitié supérieure à la médiane est également séparée en 2. Le quartile supérieur sera la valeur du point de rang 6 + 3 =9, ce qui donne Q3 = 43.

Une fois les quartiles trouvés, il est facile de mesurer la dispersion. L'écart interquartile est Q3 - Q1, ce qui donne 28 (43-15). L'écart semi-interquartile est 14 (28 ÷ 2) et l'étendue est de 43 (49-6).

Pour les ensembles de données plus grands, il est possible d'utiliser la distribution de fréquence relative cumulée pour aider à identifier les quartiles ou, encore mieux, les fonctions statistiques de base disponibles dans les tableurs et logiciels statistiques qui donnent des résultats plus aisément.

Que se passe-t-il lorsque l'ensemble de données contient un point dont la valeur est extrême par rapport au reste de la distribution?

Exemple 2 – Étendue et écart interquartile en présente d'une valeur extrême

Trouvez l'étendue et l'écart interquartile de l'ensemble de données de l'exemple 1, auquel un point de données de valeur égale à 75 est ajouté.

L'étendue sera de 69 (75-6). La médiane correspondra à la moyenne entre la valeur du point de rang n ÷ 2 = 12 ÷ 2 = 6 et celle du point de rang (n ÷ 2) + 1 = (12 ÷ 2) + 1 =7. Elle tombe donc entre le sixième et le septième rang et il y a six valeurs de chaque côté.

Le quartile inférieur sera la moyenne de la valeur du point de rang 6 ÷2 = 3 et la valeur du point de rang (6 ÷ 2) + 1 = 4. Il est donc égal à (15 + 36) ÷2 = 25,5. Le quartile supérieur sera la moyenne de la valeur du point de rang 6 + 3 = 9 et de la valeur du point de range 6 + 4 = 10, soit (43 + 47) ÷ 2 = 45. L'écart interquartile est de 45 - 25,5 = 19,5.

En résumé, l'étendue est passée de 43 à 69, une augmentation de 26 par rapport à l'exemple 1, à cause d'une seule valeur extrême. L'écart interquartile, plus robuste, est passé de 28 à 19,5, soit une diminution de 8,5 seulement.

Cet exemple permet de démontrer que l'écart interquartile est plus robuste que l'étendue lorsque l'ensemble de données contient une valeur jugée extrême. Ce n'est toutefois pas une mesure parfaite. En effet, on aurait pu s'attendre à ce que la mesure de dispersion soit un peu plus élevée en ajoutant une valeur extrême, mais le contraire s'est produit parce qu'il y avait un écart important entre les valeurs des points de rangs 3 et 4.

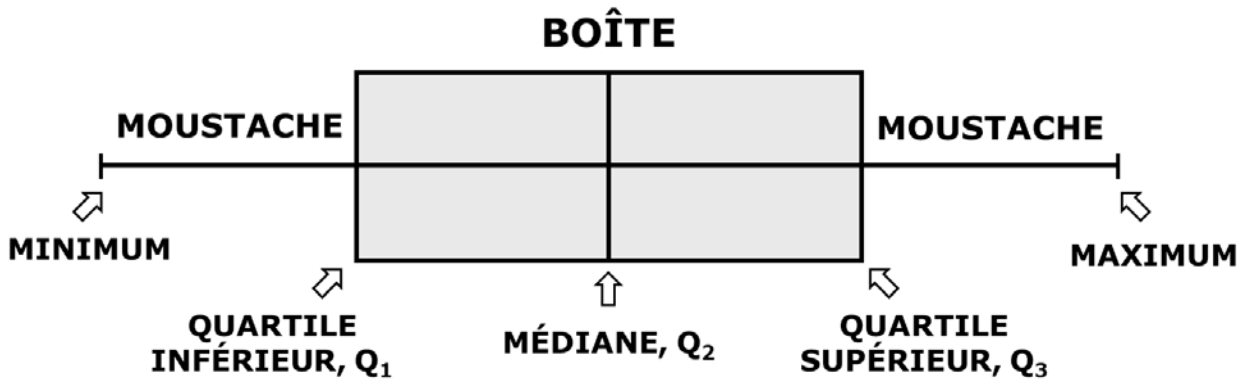
La série des cinq valeurs constituées du minimum, des trois quartiles et du maximum est désignée comme « le résumé en cinq nombres ». C'est une manière bien connue de résumer un ensemble de données. Dans la prochaine section sur la boîte à moustaches, nous verrons une méthode pratique pour visualiser le résumé en cinq nombres.

4.5.2 Visualiser la boîte à moustaches

La boîte à moustaches, parfois appelée diagramme en boîte ou diagramme de quartiles, est un type de diagramme qui permet de visualiser le résumé en cinq nombres. Elle ne montre pas la distribution avec autant de détails que l'histogramme, mais elle est particulièrement utile pour indiquer si une distribution est asymétrique et s'il y a des valeurs potentiellement extrêmes dans l'ensemble de données. La boîte à moustaches est également idéale pour comparer des distributions, car elle fait apparaître immédiatement le centre, la dispersion et l'étendue.

La figure 4.5.2.1 montre comment on construit la boîte à moustaches à partir du résumé en cinq nombres.

Figure 4.5.2.1 Construction d'une boîte à moustaches



Dans une boîte à moustaches :

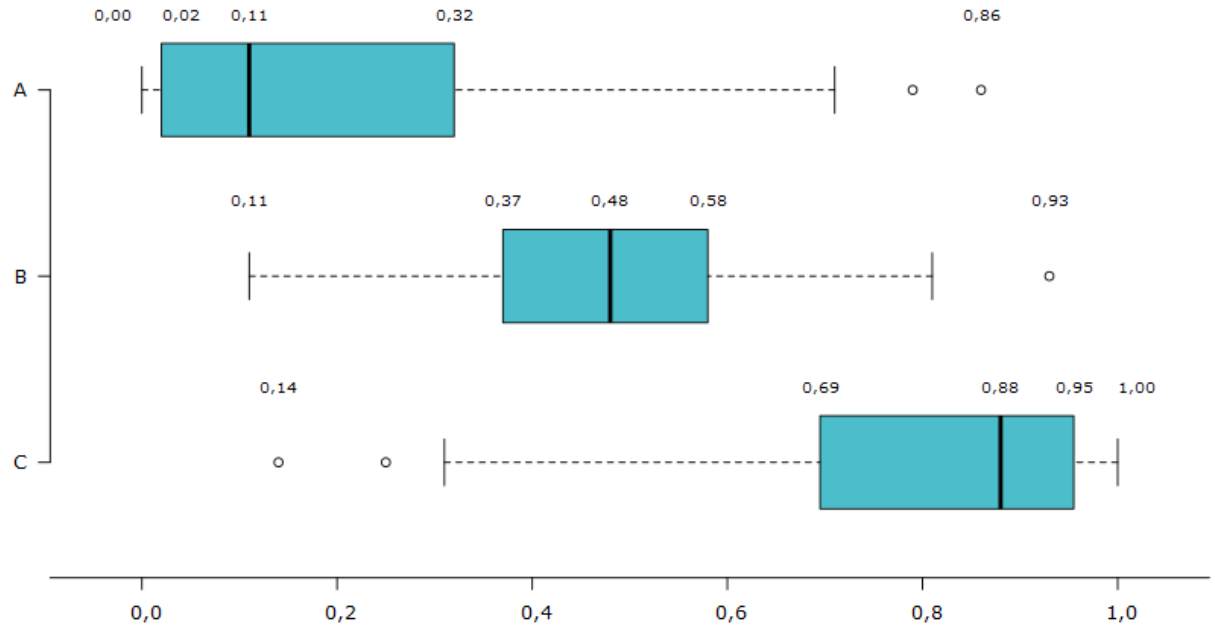
- Les côtés gauche et droit de la boîte sont les quartiles inférieur et supérieur. La boîte couvre donc l'intervalle interquartile, là où se situent 50 % des données.
- La ligne verticale qui sépare la boîte en deux représente la médiane. Parfois, la moyenne est également indiquée par un point ou une croix sur la boîte à moustaches.
- Les moustaches sont les deux lignes horizontales à l'extérieur de la boîte qui s'étendent du minimum jusqu'au quartile inférieur (le début de la boîte) et du quartile supérieur (la fin de la boîte) jusqu'au maximum.
- Le diagramme est habituellement accompagné d'un axe qui indique les valeurs (non montré à la figure 4.5.2.1).
- La boîte à moustache peut être présentée horizontalement, comme à la figure 4.5.2.1, ou verticalement.

Une variante de la boîte à moustaches restreint la longueur des moustaches à un maximum d'une fois et demi la valeur de l'écart interquartile. C'est-à-dire que la moustache s'étire jusqu'à la valeur qui est la plus éloignée du centre, mais qui respecte une distance maximale de 1,5 fois l'écart interquartile en partant du quartile inférieur ou du quartile supérieur. Les données qui dépassent cette limite sont indiquées par des points et considérées comme potentiellement extrêmes.

Exemple 1 – Comparaison de trois distributions représentées par des boîtes à moustaches

Les trois boîtes à moustaches du graphique 4.5.2.1 ci-dessous ont été créées à l'aide du logiciel R. Que peut-on affirmer à propos de ces distributions?

Graphique 4.5.2.1
Boîtes à moustaches et résumés en cinq nombres des distributions A, B et C



- Le centre de la distribution A est le plus bas des trois distributions (médiane à 0,11). La distribution est positivement asymétrique, car la portion droite de la boîte et la moustache droite sont plus longues qu'à gauche de la médiane.
- La distribution B est approximativement symétrique, car les deux moitiés de la boîte sont de longueurs sensiblement égales (0,11 du côté gauche et 0,10 du côté droit). C'est la distribution la plus concentrée, car l'écart interquartile est de 0,21, comparativement à 0,30 pour la distribution A et 0,26 pour la distribution C.
- •Le centre de la distribution C est le plus élevé des trois distributions (médiane à 0,88). La distribution C est négativement asymétrique, car la portion gauche de la boîte et la moustache gauche sont plus longues que du côté droit.

Les trois distributions incluent des valeurs potentiellement extrêmes. Prenons par exemple la distribution A. L'écart interquartile est de Q3 - Q1 = 0,32 – 0,02 = 0,30. Selon la définition utilisée par la fonction du logiciel R, toute valeur qui dépasse Q3 + 1,5 x (Q3 - Q1) = 0,32 + 1,5 x 0,30 = 0,77 se trouve à l'extérieur de la moustache et est indiquée par un cercle. Il y a deux valeurs potentiellement extrêmes dans la distribution A.

4.5.3 Calculer la variance et l'écart-type

Contrairement à l'étendue et à l'écart interquartile, la variance est une mesure qui permet de tenir compte de la dispersion de toutes les valeurs d'un ensemble de données. C'est la mesure de dispersion la plus couramment

utilisée, de même que l'écart-type, qui correspond à la racine carrée de la variance. La variance est l'écart carré moyen entre chaque donnée et le centre de la distribution représenté par la moyenne.

Exemple 1 – Calcul de la variance et de l'écart-type

Calculons la variance de l'ensemble suivant : 2, 7, 3, 12, 9.

La première étape est de calculer la moyenne. La somme est de 33 et il y a 5 nombres. La moyenne est donc de 33 ÷ 5 =6,6. Il faut ensuite calculer l'écart élevé au carré entre chaque valeur et la moyenne. Par exemple pour la première valeur :

$(2 - 6,6)^2 = 21,16$

Les écarts carrés de chaque valeur sont ensuite additionnés :

21,16 + 0,16 + 12,96 + 29,16 + 5,76 = 69,20

Cette somme est ensuite divisée par le nombre de valeurs, soit

69,20 ÷ 5 = 13,84

La variance est donc de 13,84. Il suffit de trouver la racine carrée pour obtenir l'écart-type : 3,72.

L'écart-type est utile quand on compare la dispersion de deux ensembles de données de taille semblable qui ont approximativement la même moyenne. L'étalement des valeurs autour de la moyenne est moins important dans le cas d'un ensemble de données dont l'écart-type est plus petit. Un tel ensemble renferme comparativement moins de valeurs élevées ou de valeurs faibles. Un élément sélectionné au hasard à partir d'un ensemble de données dont l'écart-type est faible peut se rapprocher davantage de la moyenne qu'un élément d'un ensemble de données dont l'écart-type est plus élevé. L'écart-type est toutefois influencé par les valeurs aberrantes. Une seule de ces valeurs pourrait avoir une grande influence sur les résultats de l'écart-type.

Il n'est pas toujours facile d'évaluer l'importance que doit avoir l'écart-type pour que les données soient largement dispersées. L'ampleur de la valeur moyenne de l'ensemble de données affecte l'interprétation de son écart-type. Lorsque vous mesurez quelque chose qui est à l'échelle de millions, avoir des mesures qui sont près de la valeur moyenne n'a pas la même signification que lorsque vous mesurez quelque chose qui est à l'échelle de centaines. Par exemple, si après avoir mesuré les recettes annuelles de deux grandes entreprises, vous constatez un écart de 10 000 $, la différence est considérée comme étant peu significative, alors que si vous mesurez le poids de deux personnes, dont l'écart est de 30 kilogrammes, la différence est considérée comme étant très significative. Voilà pourquoi il est utile, dans la plupart des cas, d'évaluer l'importance de l'écart-type par rapport à la moyenne.

Souvenez-vous des propriétés suivantes quand vous utilisez l'écart-type :

- L'écart-type est sensible aux valeurs aberrantes. Une seule valeur très aberrante peut accroître l'écart-type et, par le fait même, déformer le portrait de la dispersion.
- Pour deux ensembles de données ayant la même moyenne, celui dont l'écart-type est le plus grand est celui dans lequel les données sont les plus dispersées par rapport au centre.
- L'écart-type est égal à 0 zéro si toutes les valeurs d'un ensemble de données sont les mêmes (parce que chaque valeur est égale à la moyenne).

Ce qui explique la popularité de l'écart-type comme mesure de dispersion est son lien avec la loi normale qui décrit un grand nombre de phénomènes naturels et qui a des propriétés mathématiques intéressantes pour les grands ensembles de données. Lorsqu'une variable est distribuée selon une loi normale, l'histogramme prend la forme d'une cloche symétrique et les meilleures mesures de tendance centrale et de dispersion sont la moyenne et l'écart-type. Il s'agit d'une distribution très utile et relativement facile à utiliser. Les intervalles de confiance sont souvent basés sur la loi normale centrée réduite.

Cependant, lorsque :

- l'ensemble de données est petit,

- la distribution est asymétrique, ou
- l'ensemble de données contient des valeurs extrêmes

il est mieux d'avoir recours à l'écart interquartile.

4.6 Exercices

1. Indiquez si chacune des variables suivantes est discrète ou continue :

a. le temps qu'il faut pour vous rendre à l'école
b. le nombre de couples canadiens qui se sont mariés l'an dernier
c. le nombre de buts comptés par une équipe de hockey féminine
d. la vitesse d'une bicyclette
e. votre âge
f. le nombre de matières que vous pourrez étudier l'an prochain
g. la durée d'un appel téléphonique
h. le revenu annuel d'un particulier
i. la distance entre votre domicile et votre école
j. le nombre de pages dans un dictionnaire

2. Le propriétaire d'un dépanneur local calcule le nombre de clients qui y entrent chaque jour pendant une période de 25 jours. Voici les résultats de son calcul :

20, 21, 23, 21, 26, 24, 20, 24, 25, 22, 22, 23, 21, 24, 21, 26, 24, 22, 21, 23, 25, 22, 21, 24, 21

a. Présentez ces données au moyen d'un tableau de distribution de fréquences.
b. Quel est le résultat le plus fréquent?
c. Créez un tableau de distribution de fréquences qui inclue des colonnes pour la fréquence relative et la fréquence en pourcentage des données.

3. Quarante (40) élèves ont passé un examen de mathématiques pour lequel la note maximale qu'ils pouvaient obtenir était 10. Voici les résultats qu'ils ont obtenus :

9, 10, 7, 8, 9, 6, 5, 9, 4, 7, 1, 7, 2, 7, 8, 5, 4, 3, 10, 7, 3, 7, 8, 6, 9, 7, 4, 2, 3, 9, 4, 3, 7, 5, 5, 2, 7, 9, 7, 1

a. Construisez un tableau de fréquences de leurs notes.
b. À l'aide du tableau de fréquences, calculez la moyenne, la médiane et le mode.
c. Interprétez ces résultats.

4. Le tableau suivant fournit le nombre hypothétique de nouvelles recrues dans une grande organisation durant une période de dix ans:

Tableau 4.6.1
Nombre hypothétique de nouvelles recrues

Années	Nombre hypothétique de nouvelles recrues
1	266
2	231
3	223
4	262
5	260
6	230
7	191
8	182
9	165
10	153

a. Trouvez l'étendue.

b. Calculez l'écart interquartile.

c. Calculez le résumé en cinq nombres.

d. Dessinez un diagramme de quartiles de ces données.

4.7 Réponses

Retour à l'exercice 1a

1. a. Il s'agit d'une variable continue.
   b. Il s'agit d'une variable discrète.
   c. Il s'agit d'une variable discrète.
   d. Il s'agit d'une variable continue.
   e. L'âge exact est une variable continue. Toutefois, l'âge est souvent rapporté en arrondissant à l'entier inférieur. Dans ce cas, il s'agit d'une variable discrète.
   f. Il s'agit d'une variable discrète.
   g. Il s'agit d'une variable continue.
   h. Il s'agit d'une variable continue (peut aussi être considérée comme une variable discrète).
   i. Il s'agit d'une variable continue.
   j. Il s'agit d'une variable discrète.

2. a.

Tableau 4.7.1
Reponse pour 2a

Nombre de clients (x)	Fréquence (f)
20	2
21	7
22	4
23	3
24	5
25	2
26	2

b. L'observation la plus fréquente est 21.

c.

Tableau 4.7.2
Reponse pour 2c

Nombre de clients (X)	Fréquence (f)	Fréquence relative	Fréquence en pourcentage
20	2	0,08	8
21	7	0,28	28
22	4	0,16	16
23	3	0,12	12
24	5	0,20	20
25	2	0,08	8
26	2	0,08	8
Total	**25**	**1,00**	**100**

3. a.

Table 4.7.3
Reponse pour 3a

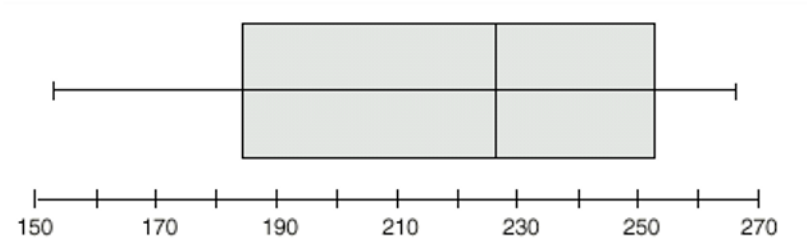
Note (x)	Fréquence (f)
1	2
2	3
3	4
4	4
5	4
6	2
7	10
8	3
9	6
10	2
Total	**40**

b. moyenne = 5,9; médiane = 7; mode = 7

c. La médiane est supérieure à la moyenne, parce que les valeurs de la plupart des observations sont élevées. La moyenne est influencée par les notes plus faibles. Le mode est égal à la médiane.

4. a. 113
   b. 78
   c. 153, 182, 226.5, 260, 266
   d.

Graphique 4.7.1
Boîte à moustaches du nombre hypothétique de nouvelles recrues



5 Visualisation des données

Les graphiques ou diagrammes sont des outils visuels efficaces. Ils présentent l'information de façon rapide et facile. Il n'est donc pas surprenant que les diagrammes soient employés fréquemment par les médias écrits et électroniques. Parfois, les données peuvent être mieux comprises lorsqu'elles sont présentées dans un graphique plutôt que dans un tableau, car le graphique peut montrer une tendance ou permettre de comparer les données.

Les élèves considèrent habituellement les diagrammes faciles à utiliser parce qu'ils sont constitués de lignes, de points et de blocs, donc des formes géométriques simples qu'ils peuvent rapidement dessiner. Dans le monde des statistiques, les graphiques montrent les relations entre des variables ou l'étendue des valeurs d'une variable ou d'un phénomène donné.

La présente section vise à décrire les graphiques de base les plus couramment utilisés pour visualiser les données. Il existe cependant un grand nombre de graphiques qui peuvent être utilisés dans différents contextes, notamment la carte thermique, la carte proportionnelle, le diagramme à bulles, le diagramme en aires, le diagramme radar ainsi que la boîte à moustaches, cette dernière ayant déjà été présentée dans une section précédente.

5.1 Utilisation des diagrammes

Lorsqu'on présente des statistiques, il est important de savoir comment illustrer l'information sous forme de graphique ou diagramme. Vous trouverez ci-dessous une liste de règles à retenir lors de la conception d'un graphique.

Un bon graphique :

- illustre les faits avec précision,
- attire l'attention du lecteur,
- complète ou démontre les arguments du texte,
- comporte un titre et des étiquettes,
- est clair et simple,
- montre les données sans modifier leur message,
- montre clairement toute tendance ou différence dans les données,
- est exact en ce qui a trait à l'aspect visuel (par exemple, pour deux valeurs, une de 15 et l'autre de 30, la deuxième devrait apparaître comme le double de la première).

Pourquoi utiliser un graphique pour présenter des données?

Les graphiques :

- se lisent et sont compris rapidement,
- montrent les faits les plus importants,
- facilitent la compréhension des données,
- peuvent convaincre le lecteur,
- aident le lecteur à se souvenir des données.

On peut utiliser différents types de graphiques pour diffuser de l'information, notamment :

- le graphique à barres,
- le pictogramme,
- le graphique circulaire,
- le graphique linéaire,
- le nuage de points,
- l'histogramme.

Il est primordial de choisir le bon type de graphique à utiliser avec un type particulier d'information. Selon la nature des données, certains graphiques sont plus appropriés que d'autres. Par exemple, il est préférable de présenter des variables catégoriques (comme les matières scolaires) sous forme de graphiques à barres ou de graphique circulaire. Par ailleurs, les variables numériques (comme la taille) sont mieux illustrées sous forme de graphique linéaire ou d'histogramme.

Graphiques : quatre directives

Si vous avez conclu qu'un graphique est le moyen de choix pour présenter de l'information, rappelez-vous ces quatre directives :

1. **Définissez le public cible**

Posez-vous les questions suivantes pour vous aider à mieux connaître votre public et ses besoins :

- Qui est le public cible?
- Que connaît-il de la question?
- Que s'attend-il à voir?
- Que veut-il savoir?
- Comment utilisera-t-il l'information?

2. **Déterminez le message à transmettre**

Posez-vous les questions suivantes pour déterminer quel est votre message et pourquoi il est important :

- Que montrent les données?
- Y a-t-il plus d'un message principal?
- Quel aspect du message doit être souligné?
- Tous les messages peuvent-ils être présentés dans un même graphique ou diagramme?

3. **Utilisez les termes appropriés pour décrire votre graphique**

Examinez le tableau suivant contenant les termes appropriés à utiliser dans le titre d'un graphique ou pour le décrire dans le texte accompagnateur :

Tableau 5.1.1 Termes pour décrire les graphiques

Si votre graphique…	Utilisez les termes suivants…
décrit des composantes	part, pourcentage de, le plus petit, la majorité
compare des éléments	rang, plus grand que, plus petit que, égal à
établit une série chronologique	changement, montée, croissance, augmentation, réduction, fluctuation
détermine une fréquence	intervalle, concentration, la plupart de, la distribution de x et y selon l'âge
analyse des relations entre les variables	augmente avec, diminue avec, varie avec, malgré, correspond à, est lié à

4. **Faites des essais avec différents types de graphique et choisissez le plus approprié**

Posez-vous les questions suivantes pour déterminer quel est votre message et pourquoi il est important :

- Graphique circulaire (description des composantes)
- Graphique à barres (comparaison des éléments et relations, série chronologique, distribution de fréquences)
- Graphique linéaire (série chronologique, distribution de fréquences)
- Nuage de points (analyse des relations)

5.2 Graphique à barres

Le graphique à barres est parfois appelé graphique à bandes ou graphique à bâtons. Il peut être horizontal ou vertical. Ce qu'il faut retenir au sujet des graphiques à barres est la longueur ou la hauteur des barres : plus elles sont longues ou hautes, plus la valeur est grande. Les graphiques à barres sont l'une des techniques employées pour présenter des données de façon visuelle pour que le lecteur puisse rapidement reconnaître un motif ou une tendance.

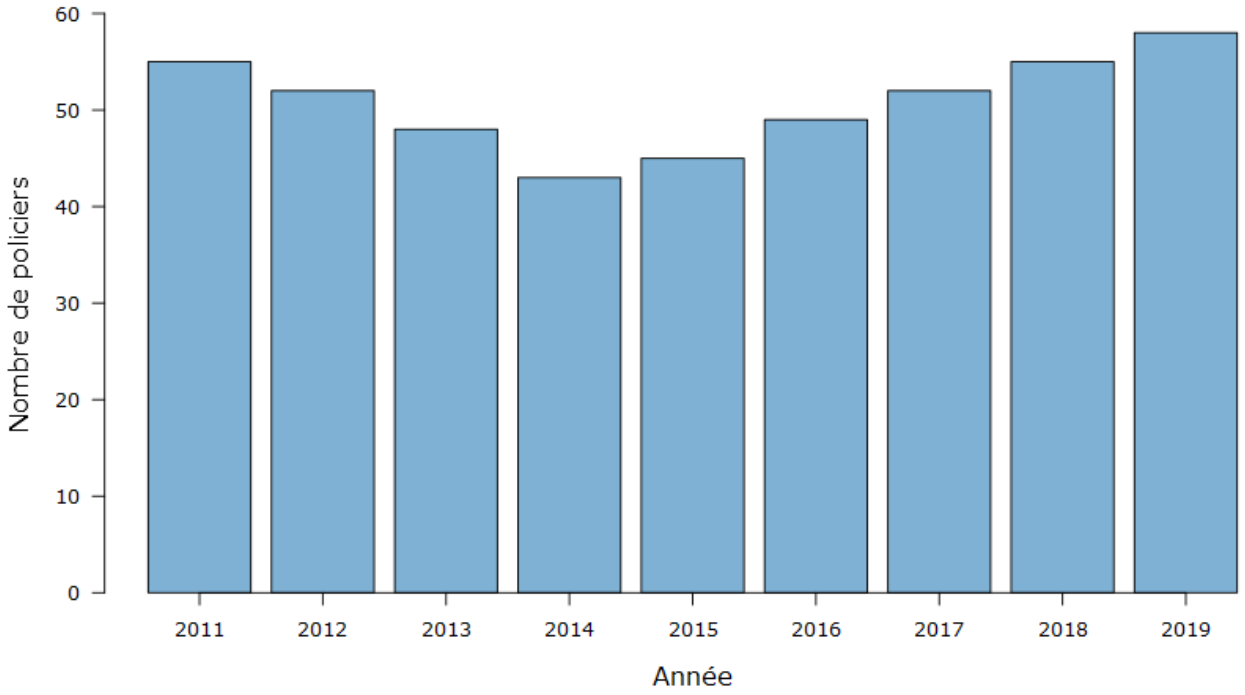
Les graphiques à barres présentent habituellement des variables catégoriques ou des variables discrètes. Ils sont composés d'un axe et d'une série de barres horizontales ou verticales. Les barres montrent les fréquences de différentes valeurs ou simplement les différentes valeurs elles-mêmes. Les nombres sur l'axe des x d'un graphique à barres horizontales ou sur l'axe des y d'un graphique à barres horizontales sont appelés l'échelle.

Lorsque vous créez un graphique à barres manuellement, dessinez une barre verticale ou horizontale pour chaque catégorie ou valeur. La hauteur ou la longueur de la barre représentera le nombre d'unités ou d'observations de cette catégorie (fréquence) ou simplement la valeur de la variable. La largeur de la barre a peu d'importance, mais elle doit être constante. Même s'il est très courant aujourd'hui qu'un logiciel, tel qu'un tableur ou le logiciel R, est utilisé pour produire des graphiques, il peut être utile de savoir comment créer des graphiques à la main.

Graphique à barres verticales

Les graphiques à barres doivent être utilisés lorsqu'on veut présenter des segments d'information. Les graphiques à barres verticales sont utiles pour comparer différentes variables catégoriques ou discrètes, comme les groupes d'âge, les classes, les écoles, etc., à condition qu'il n'y ait pas trop de catégories à comparer. Ils sont également très utiles pour les séries chronologiques. L'espace pour les étiquettes de l'axe des x est limité, mais idéal pour des années, des minutes, des heures ou des mois. Par exemple, le graphique 5.2.1 ci-dessous montre le nombre de policiers à Crimeville pour chaque année entre 2011 et 2019.

Graphique 5.2.1
Nombre de policiers à Crimeville, 2011 à 2019



Dans le graphique 5.2.1, vous pouvez voir que le nombre de policiers a diminué entre 2011 et 2014, mais qu'il a recommencé à croître en 2015. Le graphique permet également de comparer facilement le nombre de policiers entre les années.

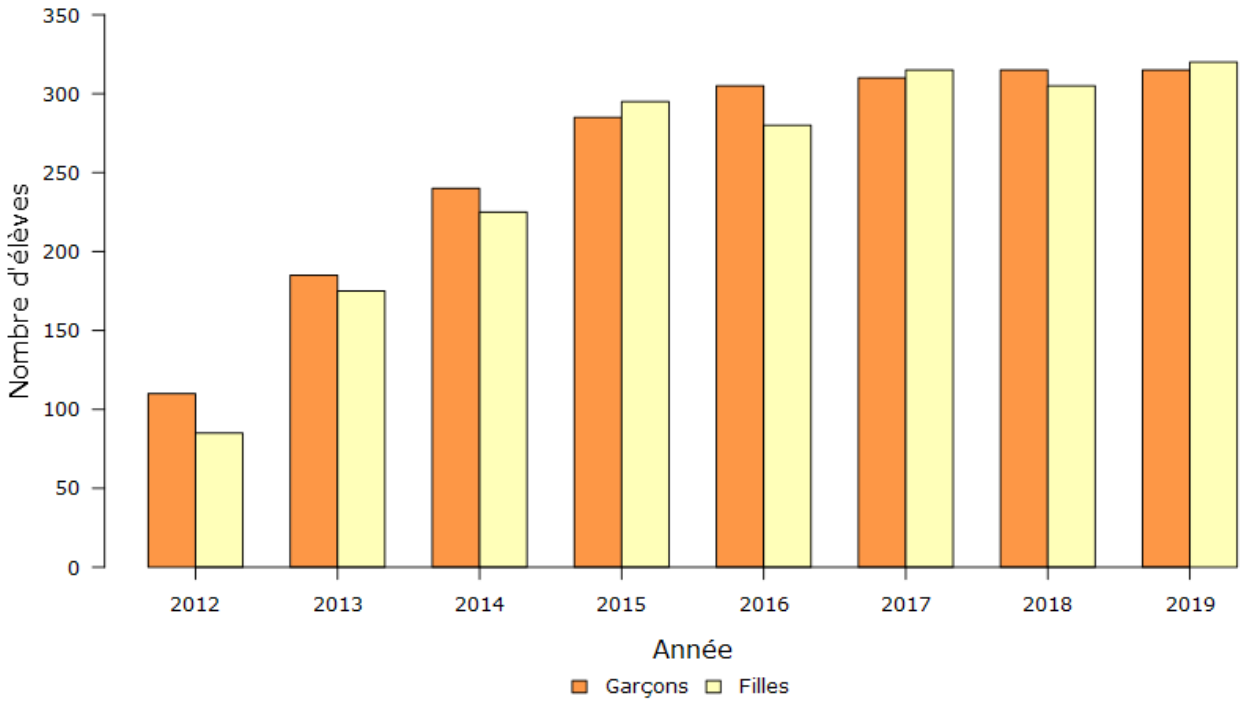
Les graphiques à barres verticales sont un excellent choix pour mettre l'accent sur un changement d'amplitude. La meilleure information à présenter dans un graphique à barres verticales est celle concernant la description des éléments, la fréquence statistique et les séries chronologiques.

Graphique à barres regroupées

Le graphique à barres regroupées est un autre moyen efficace de comparer des ensembles de données sur les mêmes endroits ou éléments. Il donne au moins deux informations pour chacun des éléments sur l'axe des x au lieu d'une seule comme dans le graphique 5.2.1. Vous pouvez faire des comparaisons directes dans un graphique selon l'âge, le genre ou tout autre élément que vous désirez comparer. Cependant, si un graphique à barres regroupées comporte trop de séries de données, le graphique devient encombré et il peut devenir difficile à lire.

Le graphique 5.2.2, un graphique à barres verticales regroupées, compare deux séries de données : le nombre de garçons et de filles qui possèdent un téléphone intelligent à l'école secondaire des Bois-Francs entre 2012 et 2019. La barre orange représente le nombre de garçons et la barre jaune, le nombre de filles.

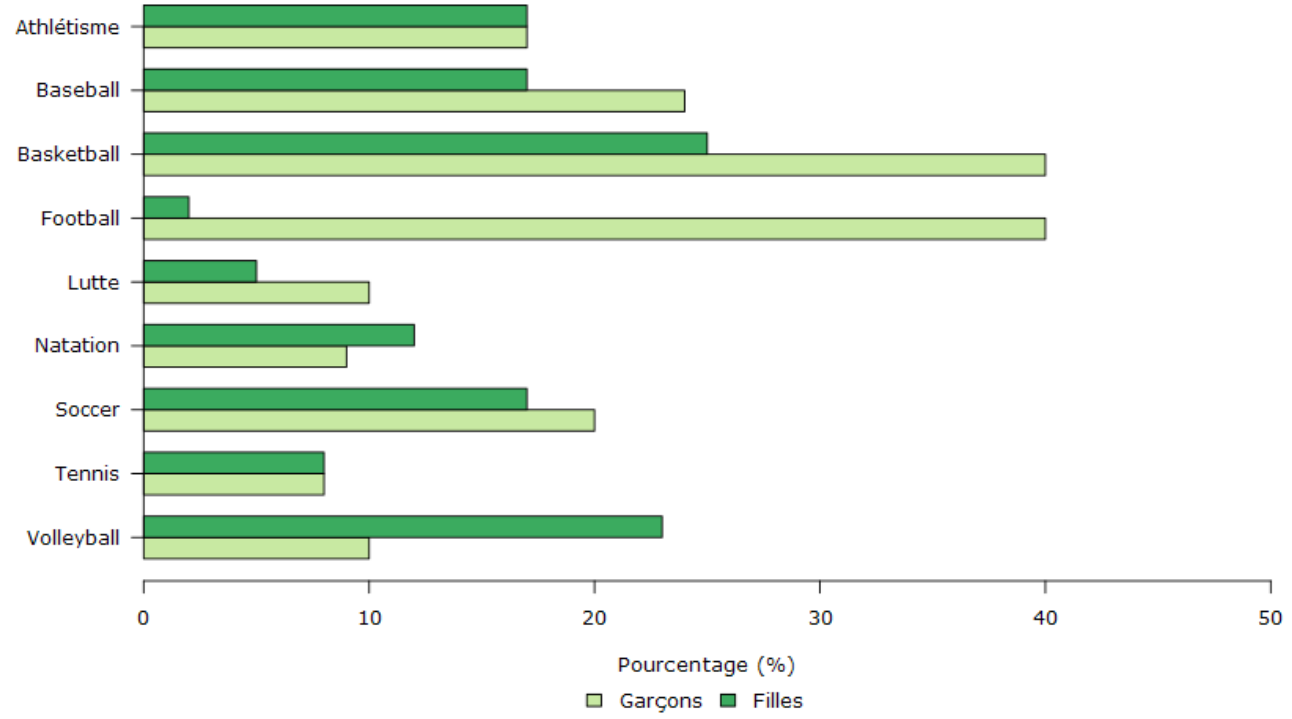
Graphique 5.2.2
Élèves possédant un télephone intelligent à l'école des Bois-Francs, selon le genre, 2012 à 2019



Graphique à barres horizontales

Un désavantage des graphiques à barres verticales, cependant, est le manque d'espace pour les étiquettes sous chacune des barres. Lorsque les étiquettes des catégories sont trop longues, il se peut qu'un graphique à barres horizontales soit préférable pour présenter l'information, comme illustré par le graphique 5.2.3.

Graphique 5.2.3
Sports pratiqués par les élèves de 15 ans de l'école de William, selon le genre



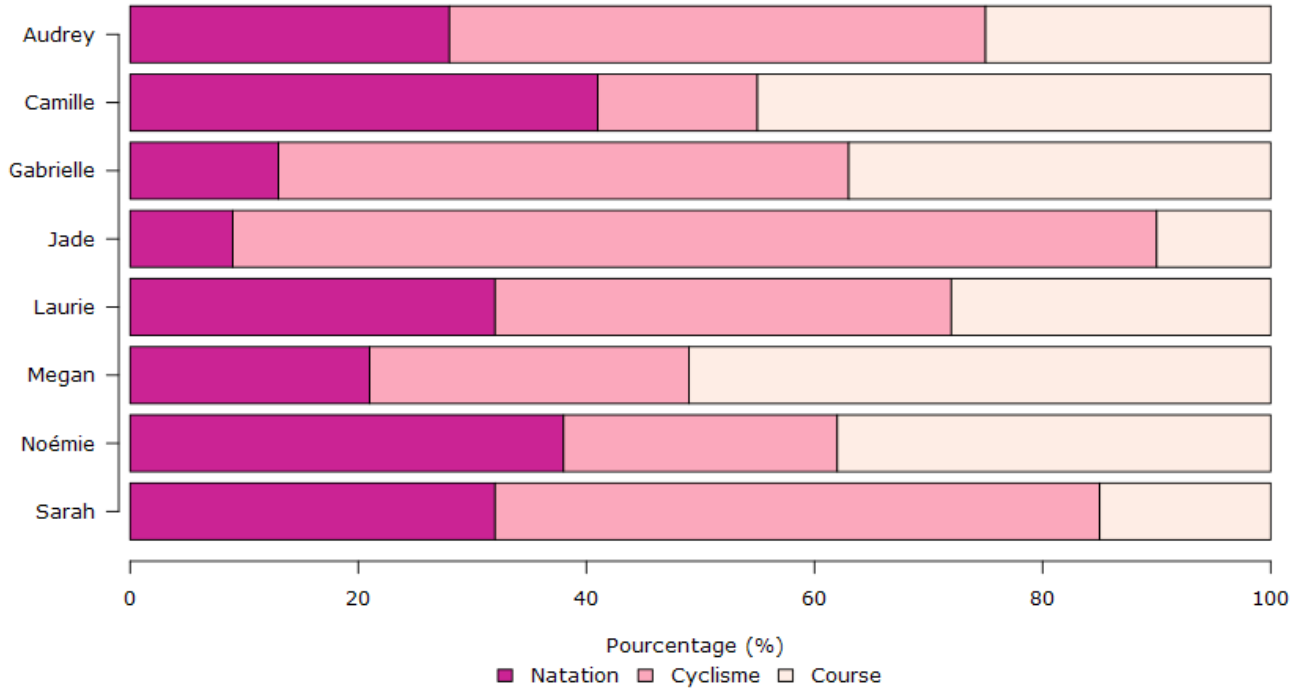
Graphique à barres empilées

Il y a plusieurs autres types de graphiques à barres qui peuvent être employés. La pyramide des âges est un exemple particulier du graphique à barres regroupées. Un autre type de graphique à barres utile est le graphique à barres empilées ou superposées.

Le graphique à barres empilées est un outil d'analyse préliminaire des données utilisé pour montrer les segments d'un tout. Le graphique à barres empilées peut être très difficile à analyser si trop d'éléments sont présentés par barre. Il peut montrer les différences entre les valeurs, mais pas nécessairement de la façon la plus simple.

Dans le graphique 5.2.4, il est facile d'analyser les données présentées puisqu'il y a seulement trois éléments par barre : natation, course et bicyclette. Il est facile de voir en un coup d'œil le pourcentage de temps consacré par chacune des femmes aux sports en question. Si ce graphique avait été utilisé pour représenter les données concernant un décathlon (10 épreuves), les données auraient été beaucoup plus difficiles à analyser.

Graphique 5.2.4
Triathlon à l'école Rousseau, pourcentage du temps consacré à chaque sport par les compétitrices



Conseils pour construire les graphiques à barres

Vous devez vous souvenir des directives suivantes lorsque vous créez des graphiques à barres :

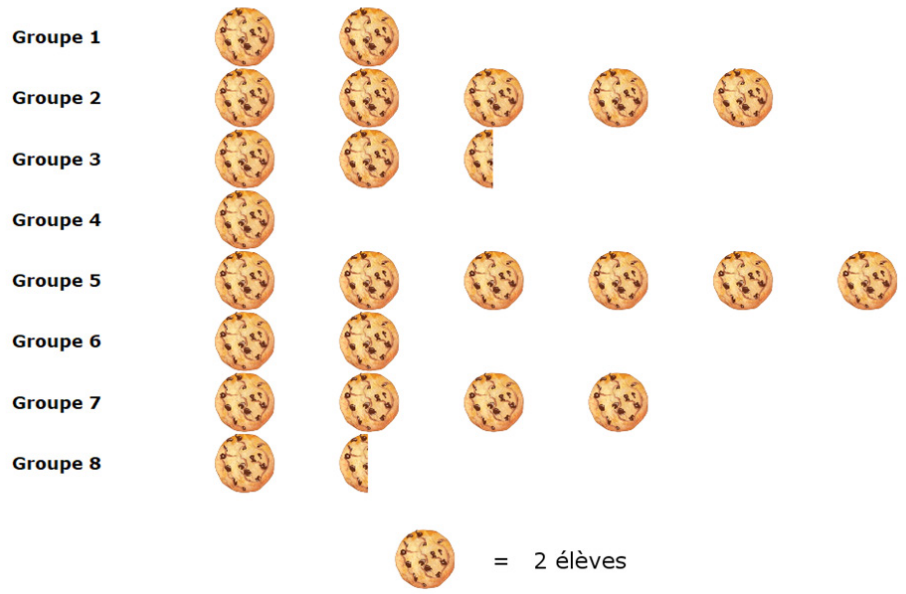
- La largeur des barres doit être plus grande que l'espace entre elles.
- N'utilisez qu'une seule police de caractère pour le texte dans le graphique. Essayez de garder la même police de caractère pour tous les graphiques d'une même présentation ou d'un même document.
- L'ordre des couleurs devrait aller de la couleur la plus foncée à la plus pâle.
- Évitez d'utiliser une combinaison de rouge et de vert dans le même affichage.

5.3 Pictogramme

Un pictogramme emploie des symboles pour illustrer l'information statistique. Il est souvent plus difficile de représenter les données avec précision à l'aide d'un pictogramme. Le pictogramme doit donc être employé avec soin pour éviter de donner une mauvaise représentation des données de façon volontaire ou involontaire.

Le graphique 5.3.1 montre le nombre d'élèves du primaire qui préfèrent les biscuits aux brisures de chocolat. Ce type de pictogramme montre comment un symbole peut être utilisé pour présenter des données. Un symbole de biscuit représente deux élèves, donc un demi-biscuit représente un élève. Ces données auraient pu facilement être présentées à l'aide d'un graphique à barres où l'échelle aurait servi à indiquer le nombre d'élèves à la place d'un symbole.

Graphique 5.3.1
Nombre d'élèves qui préfèrent les biscuits aux brisures de chocolat



Voici un autre exemple de pictogramme.

Graphique 5.3.2
Pouvoir d'achat du dollar canadien, 2000 à 2020

Le graphique 5.3.2 montre comment le dollar canadien a diminué jusqu'à atteindre une valeur de 70 cents en 20 ans en raison de l'inflation. Cette information signifie que la valeur du dollar canadien de 2020 est de 70 % la valeur du dollar de 2000! Quel est le problème avec la présentation des statistiques dans le graphique 5.3.2?

La taille des images (surface totale) des dollars (huards) dans le pictogramme peut porter à confusion. La différence entre les valeurs représentées est exagérée par les figures. Celles-ci devraient refléter le pouvoir d'achat réel des dollars pour les années visées. Puisque 70 cents est supérieur à la moitié d'un dollar, le huard pour 2020 devrait sembler plus grand que la moitié de celui de 2000, ce qui n'est pas le cas ici.

On peut prétendre que les personnes qui ont regardé le pictogramme n'ont pas été poussées à mal interpréter l'information et que ce n'est pas important. Il n'en demeure pas moins qu'inconsciemment, beaucoup de personnes feraient l'interprétation que le dollar canadien a perdu beaucoup plus de valeur qu'en réalité. Puisque beaucoup de personnes utilisent l'information statistique pour prendre des décisions, il est important qu'elle soit exacte. Dans la situation présente, la valeur décroissante du dollar canadien peut influer sur la perception qu'ont les gens de leur capacité d'économiser ou sur leur confiance en l'économie canadienne.

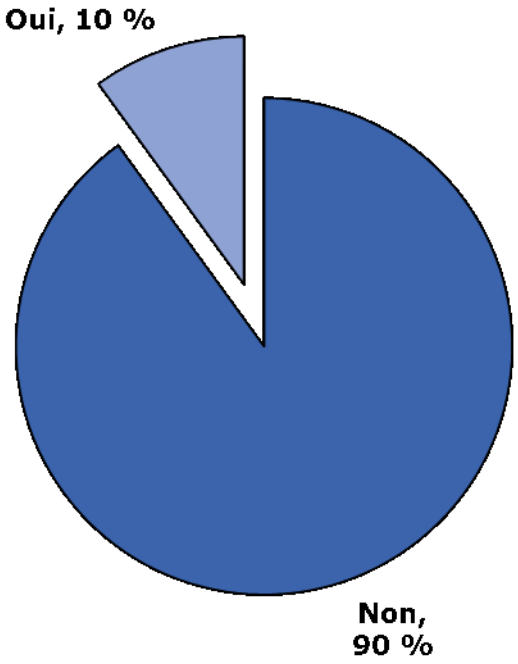
S'ils ne sont pas construits correctement, les pictogrammes peuvent être inexacts. À Statistique Canada, les pictogrammes sont rarement utilisés pour présenter l'information statistique, mais les médias les emploient très fréquemment.

5.4 Graphique circulaire

Un graphique circulaire, parfois appelé diagramme en secteurs ou camembert, est une façon de résumer un ensemble de données nominales ou de présenter les différentes valeurs d'une variable donnée (p. ex., répartition en pourcentage). Ce type de graphique est formé d'un cercle divisé en secteurs. Chaque secteur représente une catégorie particulière. La surface de chacun des secteurs représente la même proportion du cercle que la catégorie par rapport à l'ensemble des données.

Les graphiques circulaires montrent habituellement la partie d'un tout. Parfois, un secteur sera séparé du reste du cercle afin de souligner l'importance de l'information. C'est ce qu'on appelle un graphique circulaire éclaté. Le graphique 5.4.1 est exemple de graphique circulaire éclaté.

Graphique 5.4.1
Réponse des élèves et de la faculté à la question « Est-ce que les élèves de l'école Avenue devraient adopter l'uniforme? »



Le graphique 5.4.1 montre que 90 % des élèves et des membres de la faculté de l'école Avenue ne désirent pas que l'uniforme soit imposé aux élèves et que seulement 10 % des personnes de l'école le désirent. Ce fait est clairement souligné par la séparation du reste du cercle.

L'utilisation du graphique circulaire est assez répandue, puisque le cercle représente le concept d'ensemble (100 %). Les graphiques circulaires sont également parmi les plus utilisés en raison de leur facilité d'emploi. Bien qu'ils soient souvent employés, les graphiques circulaires doivent être utilisés soigneusement pour deux raisons. Premièrement, ils sont utiles pour présenter l'information lorsqu'il n'y a qu'un maximum de cinq ou six éléments. S'il y a davantage d'éléments, la figure créée sera trop difficile à comprendre. Deuxièmement, les graphiques circulaires ne sont pas utiles lorsque les valeurs des composantes sont trop semblables parce qu'il peut être difficile de voir les différences de taille.

Le graphique circulaire utilise les pourcentages pour comparer les catégories. Les pourcentages sont utilisés parce qu'il est plus facile de représenter un tout de cette façon. Le tout est égal à 100 %. Par exemple, si vous passez sept heures à l'école et que 55 minutes de ce temps sont consacrées à votre dîner, alors 13,1 % de votre journée à l'école est consacrée au dîner. Pour présenter cette information dans un graphique circulaire, vous devriez trouver combien de degrés représentent 13,1 %. Ce calcul est effectué grâce à l'équation suivante :

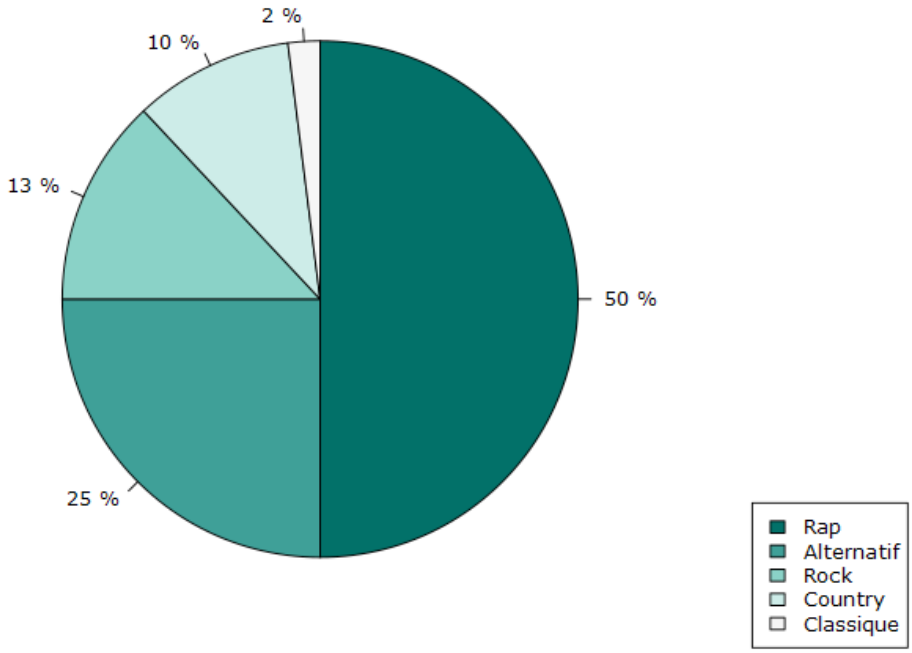
pourcentage ÷ 100 x 360 degrés = nombre de degrés

Le ratio fonctionne parce que le pourcentage total du cercle représente 100 % et qu'il y a 360 degrés dans un cercle. C'est donc 47,1 degrés du cercle (13,1 %) qui correspond au temps consacré au dîner.

Construction d'un graphique circulaire

Un graphique circulaire est construit en convertissant la part de chacune des séries en un pourcentage de 360 degrés. Dans le graphique 5.4.2, on peut bien voir les préférences musicales des 14 à 19 ans.

Graphique 5.4.2
Genres musicaux préférés des jeunes adultes de 14 à 19 ans



Le graphique circulaire vous indique rapidement les éléments suivants :

- La moitié des élèves préfèrent le rap (50 %),
- Les autres élèves préfèrent la musique alternative (25 %), le rock and roll (13 %), le country (10 %) et la musique classique (2 %).

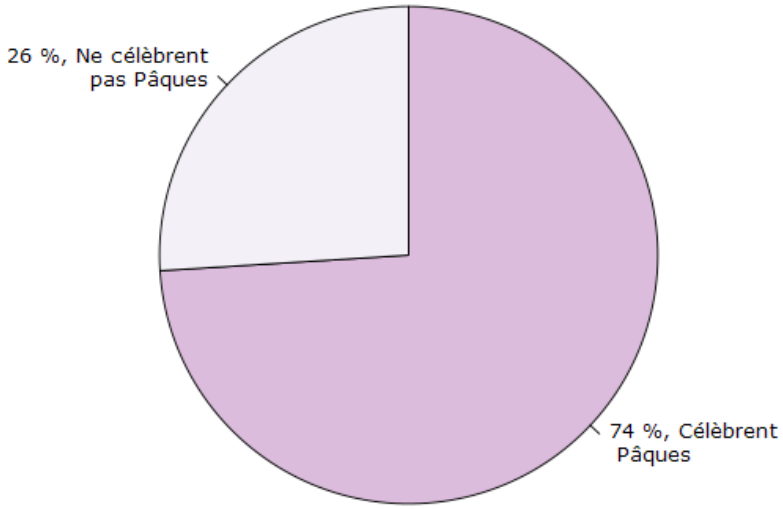
Astuce! Lorsque vous créez un graphique circulaire, assurez-vous que les secteurs sont en ordre de grandeur (du plus grand au plus petit) dans le sens horaire.

Afin de reproduire le graphique circulaire ci-dessus, suivez les étapes suivantes :
Si 50 % des élèves préfèrent le rap, 50 % du cercle (360 degrés) équivaudraient à 180 degrés.

- Dessinez un cercle avec votre rapporteur d'angles.
- À partir du sommet du cercle, mesurez un angle de 180 degrés avec votre rapporteur. L'élément « rap » devrait compter pour la moitié du cercle. Placez une indication à cet endroit avec votre règle.
- Reprenez le processus pour chacune des catégories musicales, en dessinant un angle correspondant au pourcentage de 360 degrés. La catégorie finale ne doit pas être mesurée puisque son rayon est déjà en place.

Placer une étiquette de pourcentage peut faciliter la lecture du graphique. S'il y a peu de catégories, le pourcentage et l'étiquette de la catégorie devraient être placés à côté du secteur, comme au graphique 5.4.3. De cette façon, les utilisateurs n'ont pas à revenir constamment à la légende pour savoir quelle catégorie est représentée par telle couleur.

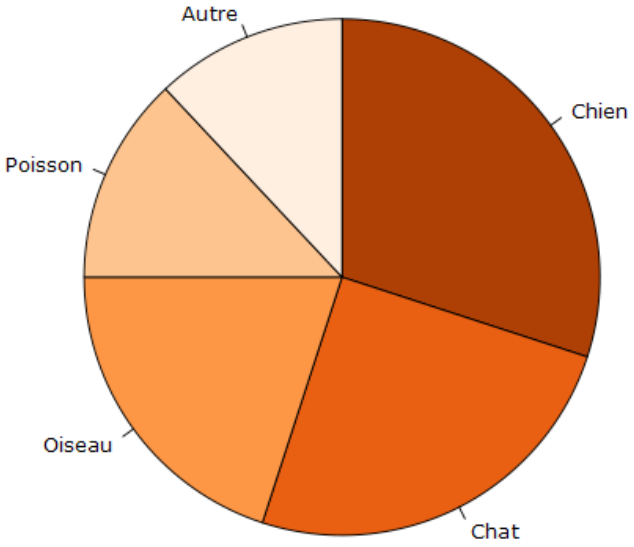
Graphique 5.4.3
Pourcentage des élèves de la classe des religions du monde de M. Paul qui célèbrent Pâques



Le graphique 5.4.3 transmet un message clair à l'utilisateur : 74 % des élèves de la classe de religions du monde célèbrent Pâques. On peut facilement comprendre le message simplement en regardant les pourcentages.

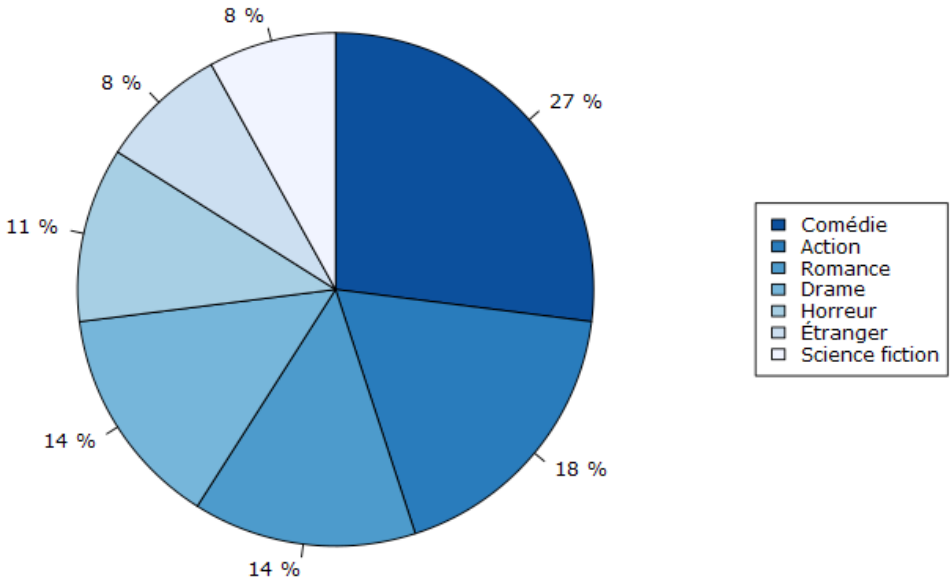
Il est plus difficile de comprendre le message lié au graphique 5.4.4 parce qu'il n'y a pas de pourcentage lié aux secteurs. L'utilisateur peut quand même se faire une idée de ce que l'on veut dire au sujet des animaux de compagnie vendus dans le magasin, mais le message n'est pas aussi clair que si les secteurs du graphique avaient été accompagnés d'une étiquette.

Graphique 5.4.4
Animaux de compagnie achetés chez le Monde des animaux



Dans le graphique 5.4.5 ci-dessous, la légende est bien construite et les pourcentages accompagnent les secteurs. Cependant, il y a trop d'éléments dans le graphique pour qu'il soit possible d'avoir une idée rapide de la répartition des préférences. S'il y a plus de cinq ou six catégories, songez à utiliser un autre type de graphique pour présenter l'information. Dans le graphique 5.4.5, il aurait certainement été préférable d'utiliser un graphique à barres.

Graphique 5.4.5
Types de film préférés dans la classe de cinéma de Mme Robert

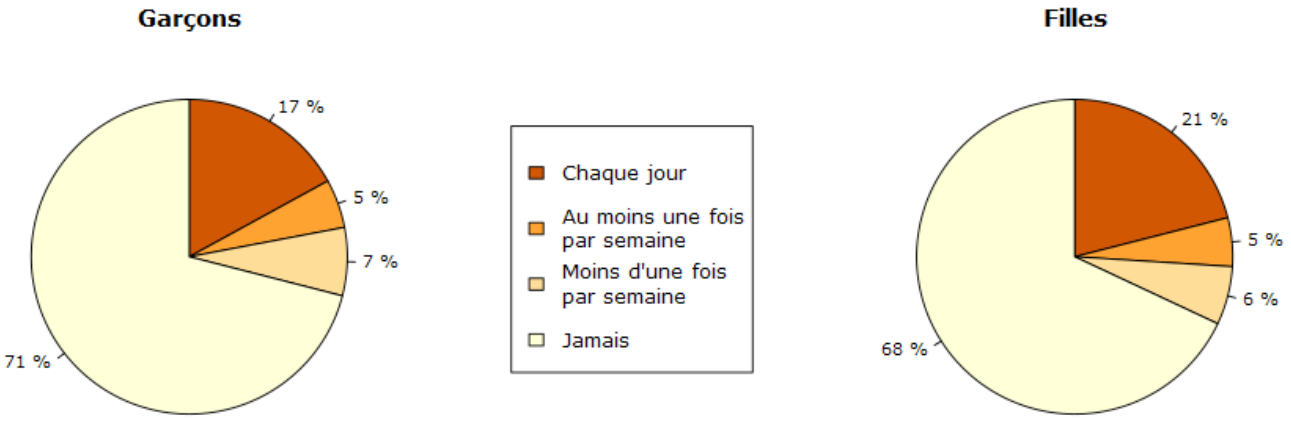


Astuce! De nombreux logiciels, comme les tableurs, peuvent dessiner les graphiques circulaires rapidement et facilement. Cependant, les recherches ont montré que beaucoup de gens pouvaient se tromper en lisant un graphique circulaire. En règle générale, les graphiques à barres peuvent communiquer le même message et ils risquent moins de créer une confusion.

Graphiques circulaires et graphiques à barres

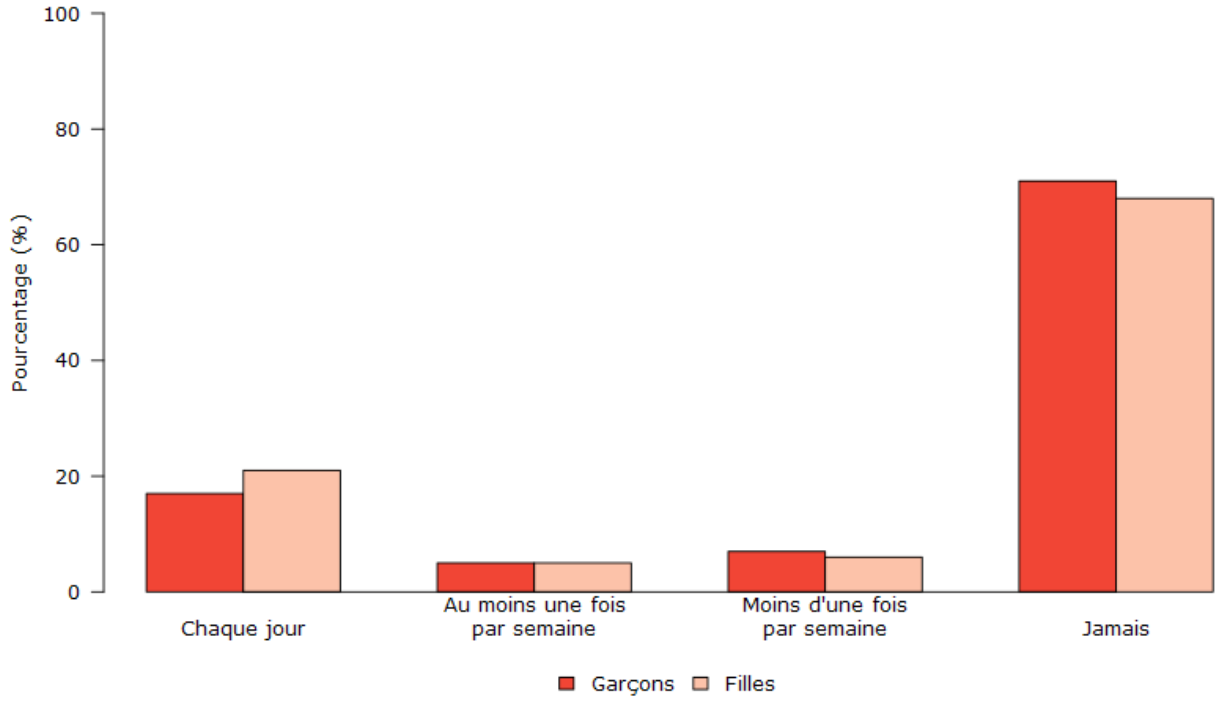
Lorsque vous présentez de l'information statistique, n'utilisez pas plus d'un graphique circulaire. Le graphique 5.4.6 montre deux graphiques circulaires côte à côte, alors qu'un graphique à barres regroupées aurait mieux présenté l'information. Un utilisateur pourrait éprouver des difficultés à comparer les secteurs d'un graphique à ceux de l'autre graphique. Par contre, dans un graphique à barres regroupées, les secteurs deviennent des barres mises l'une à côté de l'autre, ce qui facilite la comparaison.

Graphique 5.4.6
Tabagisme chez les membres de l'équipe d'athlétisme de 15 ans de l'école secondaire du Parc, selon le genre



Le graphique 5.4.7 montre comment un graphique à barres regroupées serait un meilleur choix pour présenter l'information comparativement à deux graphiques circulaires. L'important pour la préparation de ce type de graphique est de s'assurer d'utiliser la même échelle pour les deux barres du diagramme. Vous verrez que l'information est beaucoup plus claire dans le graphique 5.4.7 que dans le graphique 5.4.6.

Graphique 5.4.7
Tabagisme chez les membres de l'équipe d'athlétisme de 15 ans de l'école secondaire du Parc, selon le genre



5.5 Graphique linéaire

Les graphiques linéaires, surtout employés en statistique et en science, sont utilisés plus fréquemment que tous les autres types de graphiques, parce que leurs caractéristiques visuelles révèlent clairement les tendances dans les données de façon claire et qu'il s'agit d'un type de graphique facile à créer.

Un graphique linéaire est une comparaison visuelle des relations entre deux variables placées sur l'axe des x et l'axe des y. Il montre les liens entre les informations en plaçant un trait continu entre les points d'une grille.

Les graphiques linéaires comparent deux variables : l'une est placée sur l'axe des x (horizontal) et l'autre sur l'axe des y (vertical). L'axe des y dans un graphique linéaire indique habituellement une quantité (p. ex., dollars, litres) ou un pourcentage, alors que l'axe des x horizontal sert souvent à mesurer les unités de temps. C'est pourquoi le graphique linéaire est souvent considéré comme un graphique de séries chronologiques. Par exemple, si vous vouliez montrer dans un graphique la hauteur d'un lancer de baseball, vous pourriez placer la variable du temps sur l'axe des x et la hauteur sur l'axe des y. Bien qu'ils ne présentent pas les données spécifiques aussi bien que les tableaux, les graphiques linéaires peuvent montrer les relations plus clairement que les tableaux. Ils peuvent également montrer de nombreuses séries et conviennent donc habituellement mieux aux séries chronologiques et aux distributions statistiques.

Les graphiques à barres verticales et les graphiques linéaires ont un objectif commun. Le graphique à barres verticales présente un changement d'amplitude, alors que le graphique linéaire est employé pour montrer un changement d'orientation.

En résumé, les graphiques linéaires :

- montrent bien les valeurs de données spécifiques,
- révèlent les tendances et les relations entre les données,
- comparent les tendances de différents groupes.

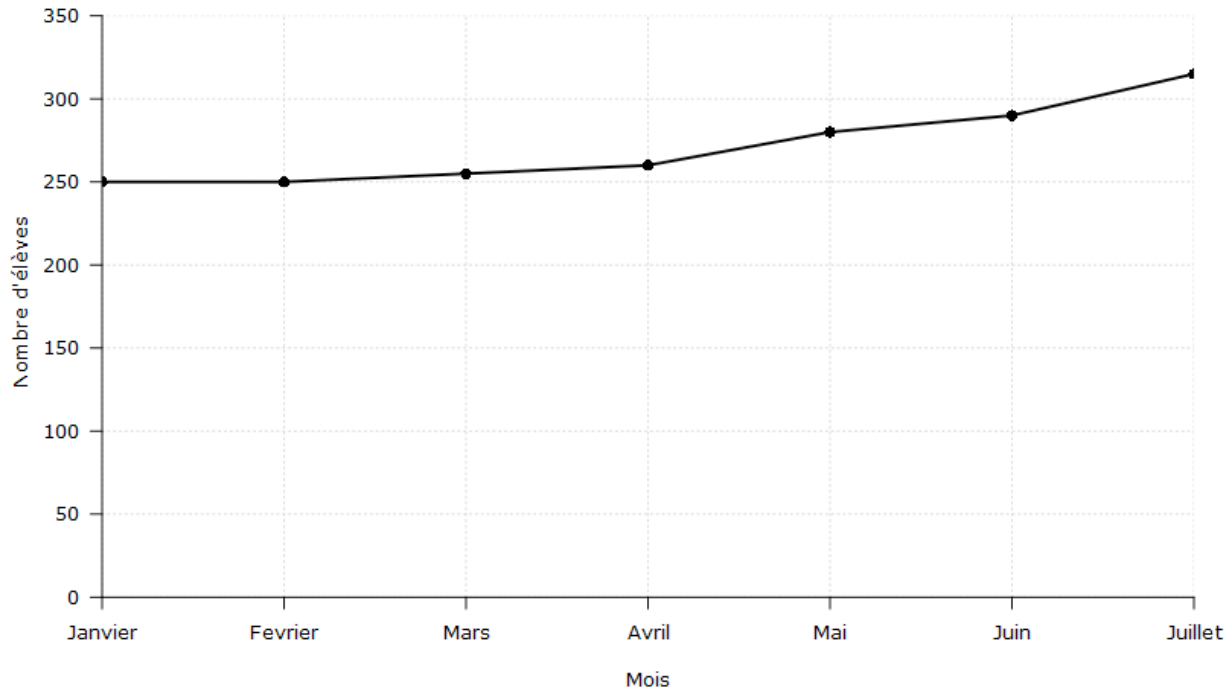
Les graphiques peuvent parfois donner une représentation déformée des données. Si les échelles employées pour les axes d'un graphique font paraître les données d'une certaine façon, alors le graphique peut faire ressortir une tendance autre que la tendance réelle. C'est le cas si les intervalles entre les points adjacents d'un axe sont inégaux ou si la même donnée, placée dans deux échelles différentes, semble différente.

Exemple 1 – Extraire une tendance temporelle

Le graphique 5.5.1 montre une tendance évidente, la variation de la population active de janvier à juillet. Le nombre d'élèves de l'école d'Olivier qui font partie de la population active est présenté sur l'axe des y (l'axe vertical), alors que la variable du temps se trouve à l'axe des x (l'axe horizontal).

Le nombre d'élèves qui font partie de la population active était de 252 en janvier, 252 en février, 255 en mars, 256 en avril, 282 en mai, 290 en juin et 319 en juillet. En examinant le graphique plus attentivement, on s'aperçoit que la participation des élèves était constante pour les quatre premiers mois (janvier à avril) et que pour les trois autres mois (mai à juillet) le nombre s'est accru de façon constante.

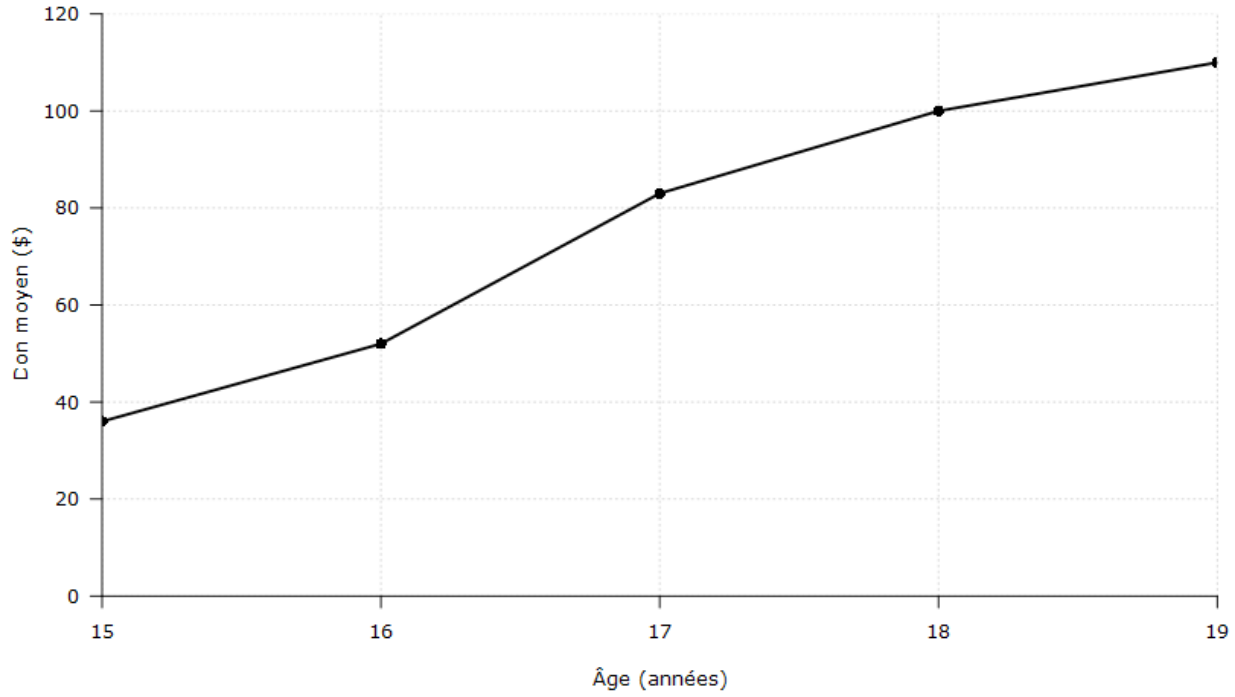
Graphique 5.5.1
Participation à la population active à l'école d'Olivier



Exemple 2 – Comparer deux variables liées

Le graphique 5.5.2 est un graphique linéaire simple qui compare deux éléments. Dans cet exemple, le temps n'est pas un facteur. Le graphique compare le nombre moyen de dollars donnés selon l'âge des donneurs. Selon la tendance dans le graphique, plus le donneur est âgé, plus il donne un montant important. Le donneur de 17 ans donne, en moyenne, 84 $. Chez les donneurs âgés de 19 ans, le don moyen est supérieur de 26 $, soit 110 $.

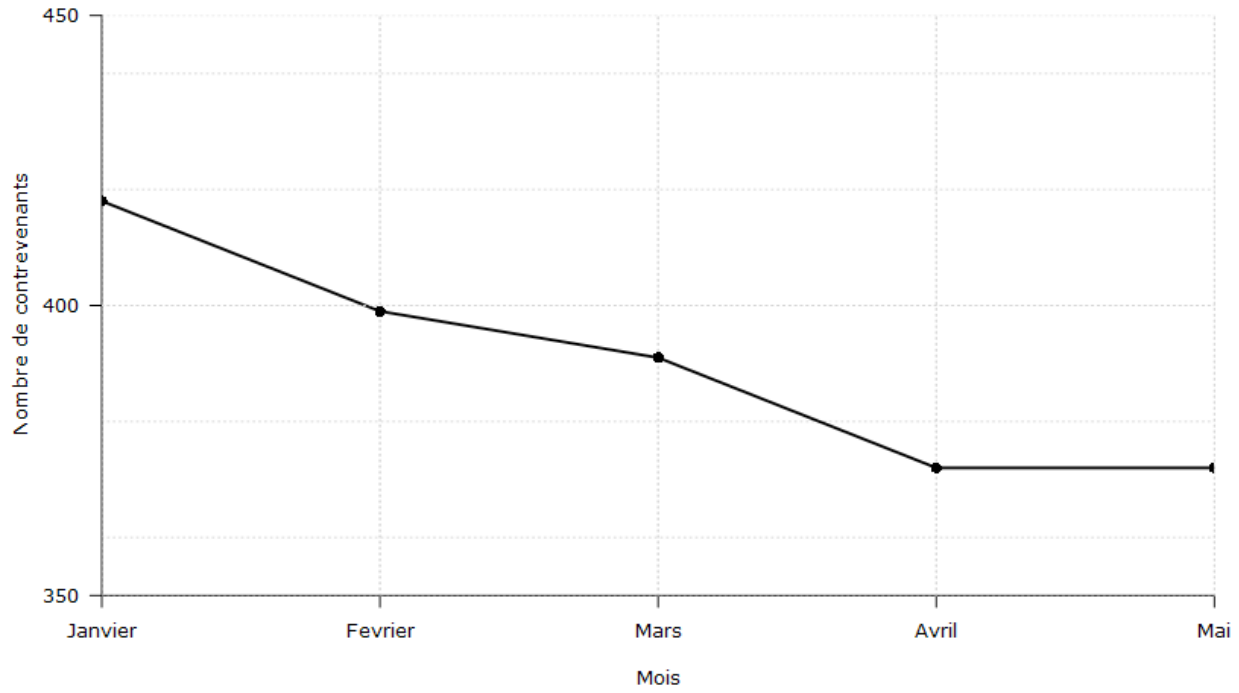
Graphique 5.5.2
Nombre moyen de dollars donnés à l'école Bois-Verts, selon l'âge des donateurs



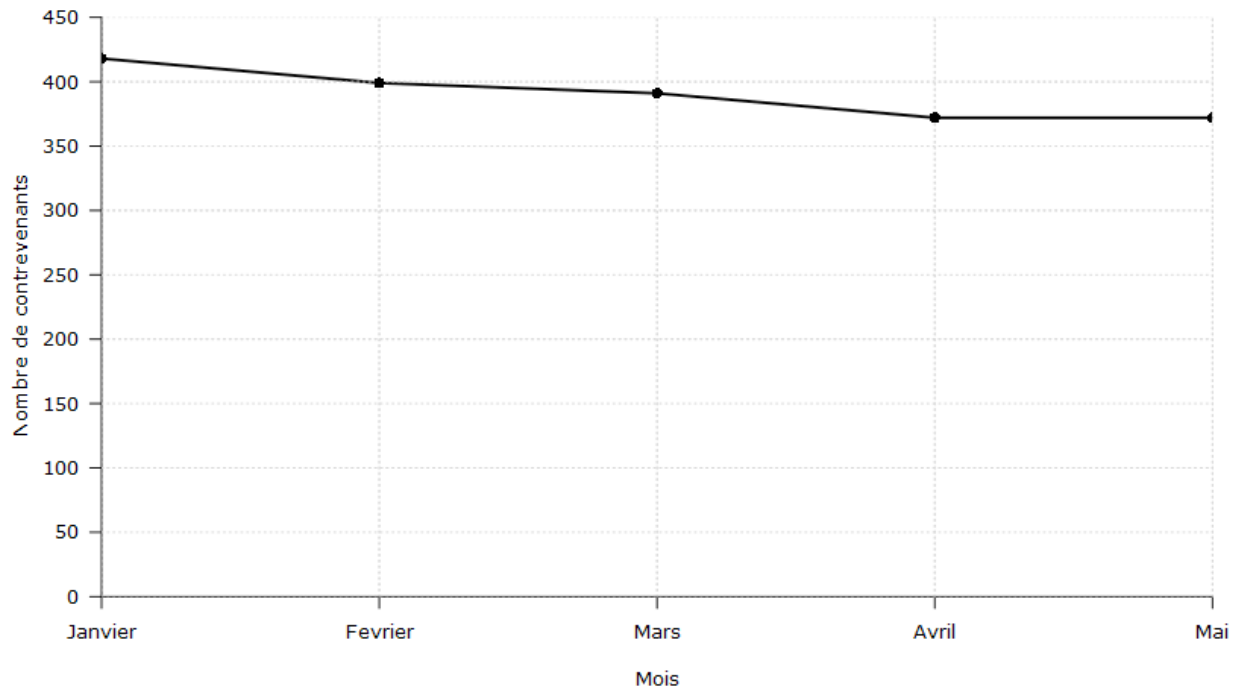
Exemple 3 – Utiliser la bonne échelle

Lorsque vous tracez un axe, il est important d'utiliser la bonne échelle. Si vous ne le faites pas, l'axe pourra donner au lecteur une impression fausse quant aux données. Comparez le graphique 5.5.3 et le graphique 5.5.4 :

Graphique 5.5.3
Nombre de contrevenants reconnus coupables à Grishamville



Graphique 5.5.4
Nombre de contrevenants reconnus coupables à Grishamville



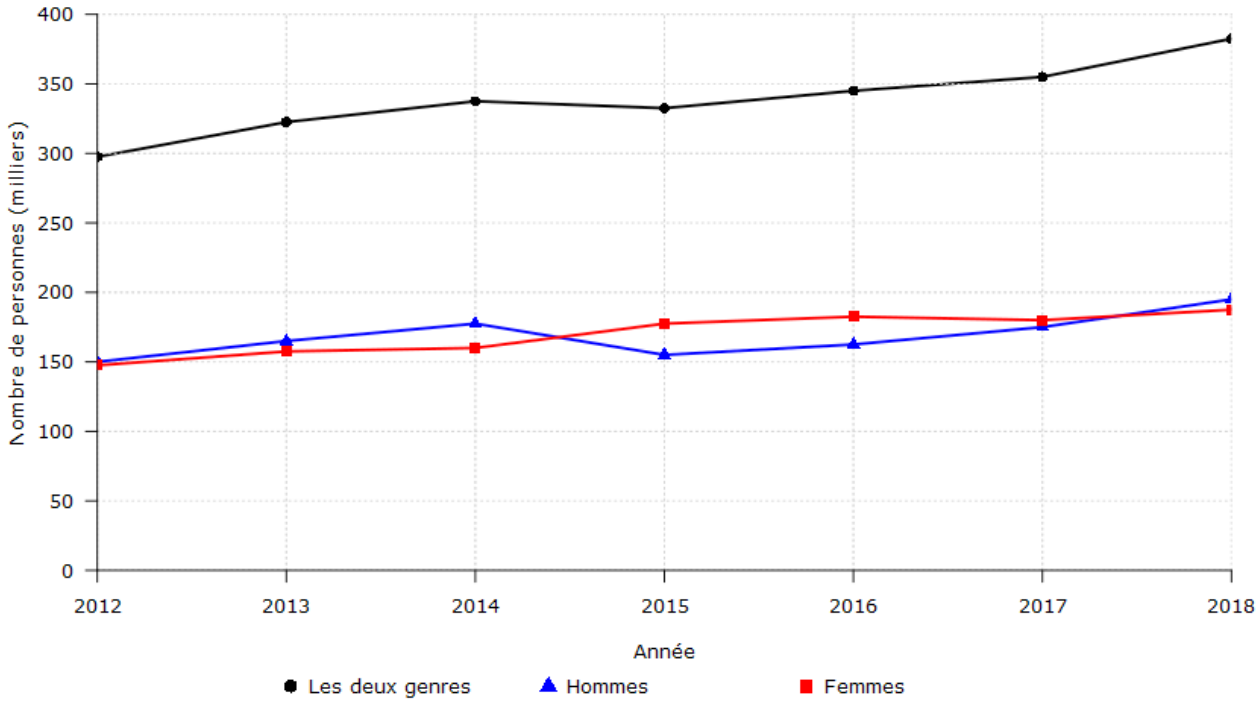
L'utilisation d'une échelle de 350 à 430 (graphique 5.5.3) met l'accent sur un petit éventail de valeurs. Elle ne permet pas de montrer clairement la tendance en ce qui a trait aux contrevenants reconnus coupables entre janvier et mai, puisqu'elle exagère la tendance. Cependant, le choix d'une échelle de 0 à 450 (graphique 5.5.4) montre mieux à quel point le déclin des contrevenants reconnus coupables est réellement minime.

Les deux graphiques peuvent avoir leur utilité selon le contexte. La chose importante à retenir, c'est qu'il faut toujours porter attention à l'échelle lors de l'interprétation d'un graphique.

Exemple 4 – Graphiques linéaires multiples

Les graphiques linéaires multiples permettent de comparer efficacement des éléments semblables pour une même période, comme le montre le graphique 5.5.5 qui compare l'utilisation du téléphone cellulaire selon le genre.

Graphique 5.5.5
Utilisation des téléphones cellulaires par genre, Touteville, 2012 à 2018



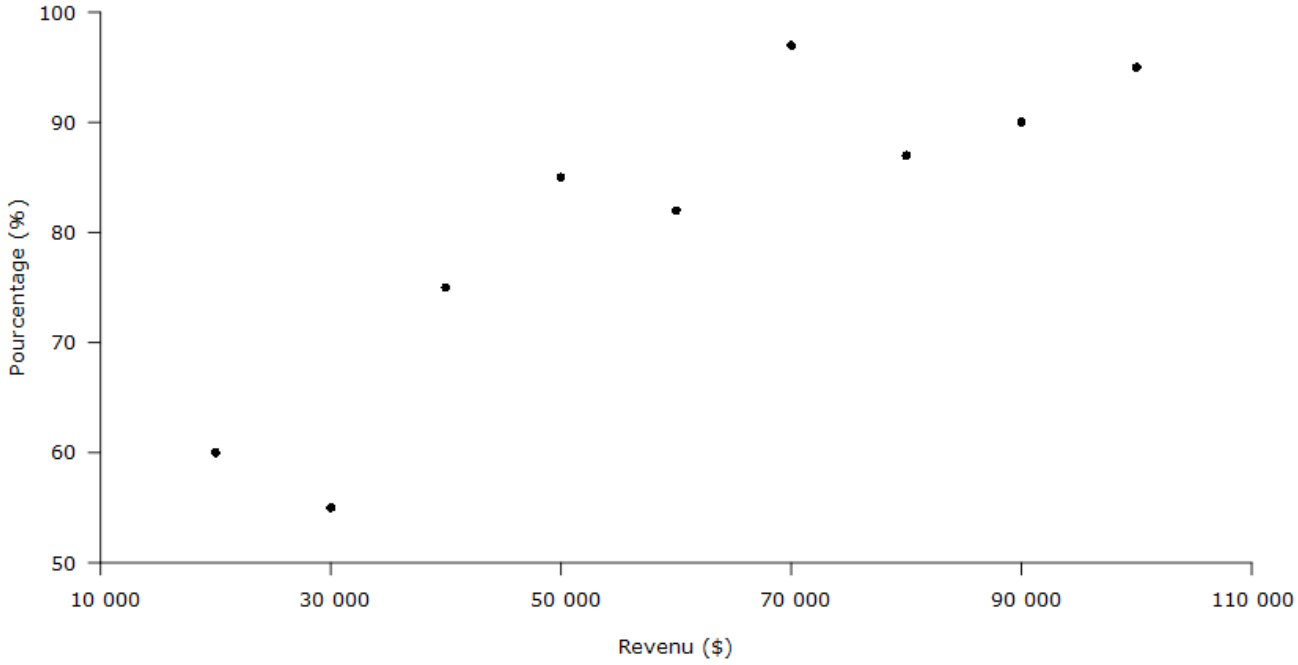
Le graphique 5.5.5 est un exemple d'un bon graphique. Le message est énoncé clairement dans le titre et chaque ligne du diagramme est identifiée correctement. Il est facile de voir dans ce graphique que l'utilisation totale des téléphones cellulaires s'est accrue de façon constante depuis 2012, sauf pendant une période d'un an (2015) pendant laquelle le nombre a diminué quelque peu. La tendance d'utilisation pour les femmes et les hommes semble être très semblable, malgré quelques différences mineures.

5.6 Nuage de points

En science, le nuage de points est grandement utilisé pour présenter la mesure de deux ou plusieurs variables liées. Le nuage de points est particulièrement utile lorsque les valeurs des variables sur l'axe des y dépendent des valeurs de la variable de l'axe des x.

Dans un nuage de points, les points sont placés sans être reliés. La tendance qui en résulte indique le type et la force de la relation entre deux ou plusieurs variables. Le graphique 5.6.1 est un exemple de nuage de points. Le pourcentage de gens qui possèdent une voiture augmente avec le revenu, ce qui montre une relation positive entre ces deux variables.

Graphique 5.6.1
Possession d'une voiture à Touteville, selon le revenu du ménage



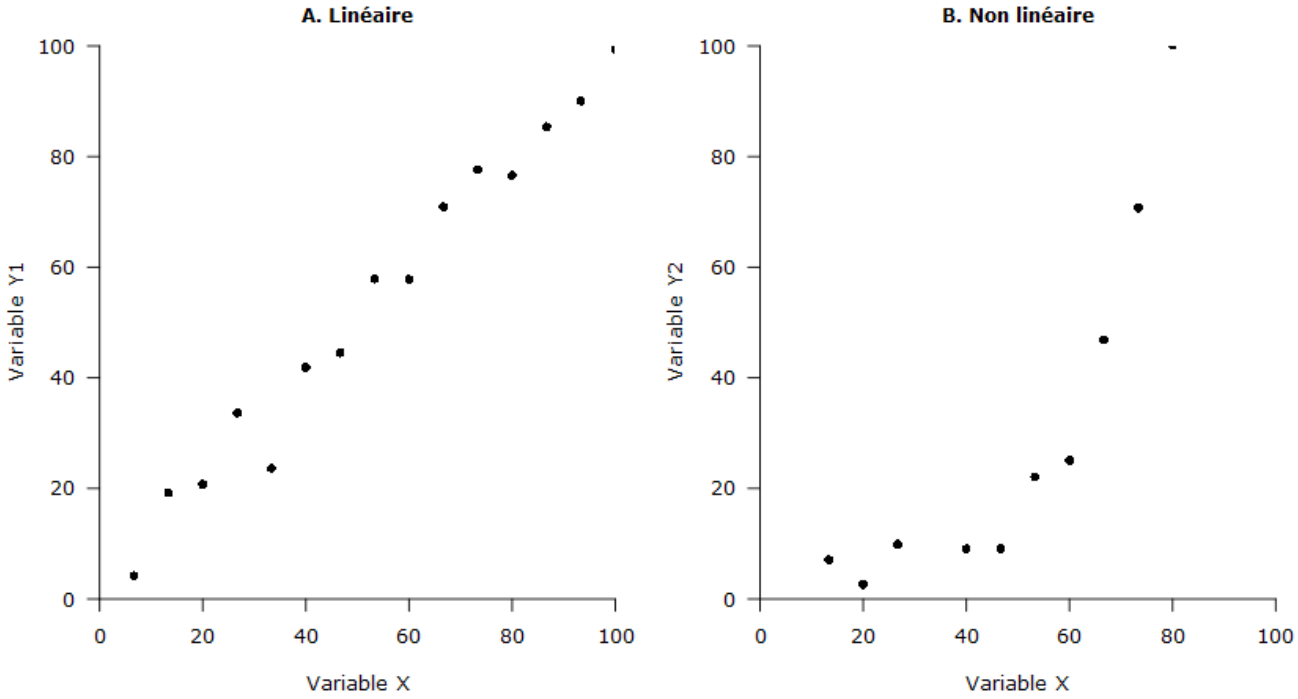
La tendance des points du nuage montre la relation entre les variables. Les nuages de points peuvent montrer différentes tendances et relations, par exemple :

- une relation linéaire ou non linéaire,
- une relation positive (directe) ou négative (inverse),
- la concentration ou la dispersion des données,
- la présence de valeurs extrêmes.

Relation linéaire ou non linéaire

Lorsque les points forment une ligne droite dans le graphique, la relation entre les variables est linéaire, comme dans le graphique 5.6.2, partie A. Lorsque les points ne forment pas de ligne ou forment une ligne qui n'est pas droite comme au graphique 5.6.2, partie B, la relation n'est pas linéaire.

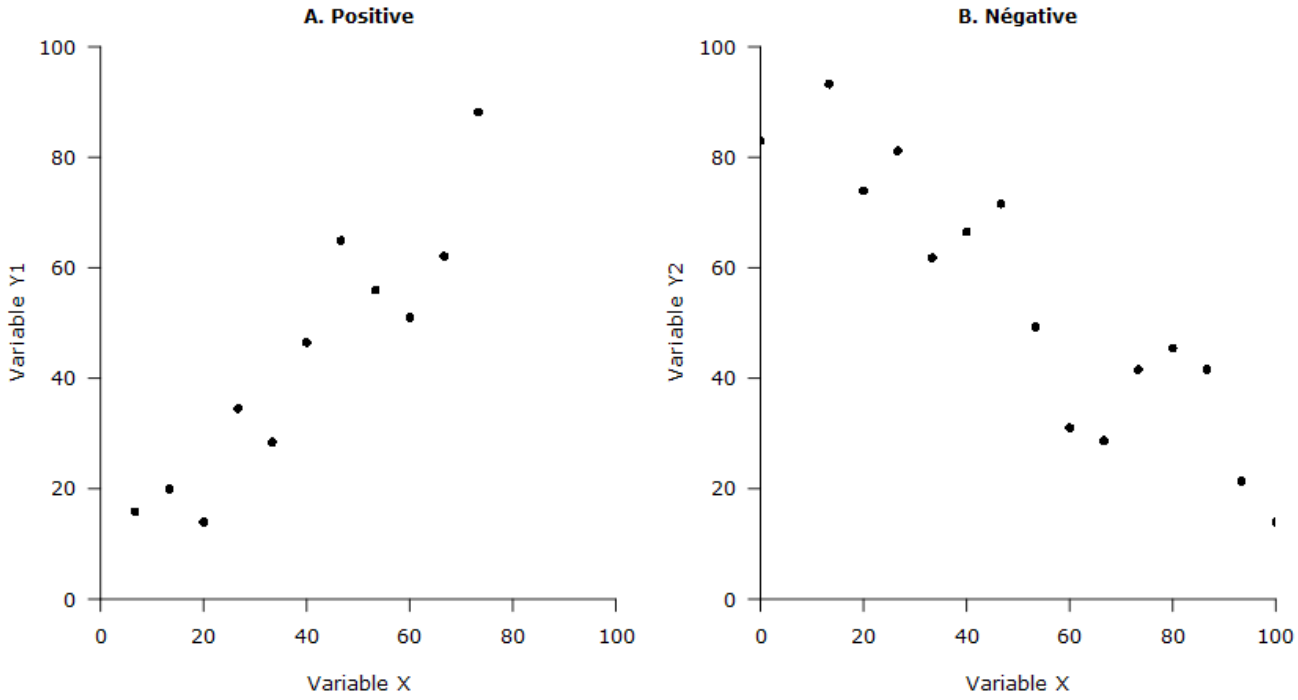
Graphique 5.6.2
Relation linéaire ou relation non linéaire



Relation positive ou négative

Si les points sont regroupés près d'une ligne qui va du coin inférieur gauche au coin supérieur droit du graphique, la relation entre les deux variables est dite positive ou directe (graphique 5.6.3, partie A). Si les points sont regroupés près d'une ligne qui va du coin supérieur gauche au coin inférieur droit du graphique, la relation entre les variables est dite négative ou inverse (graphique 5.6.3, partie B).

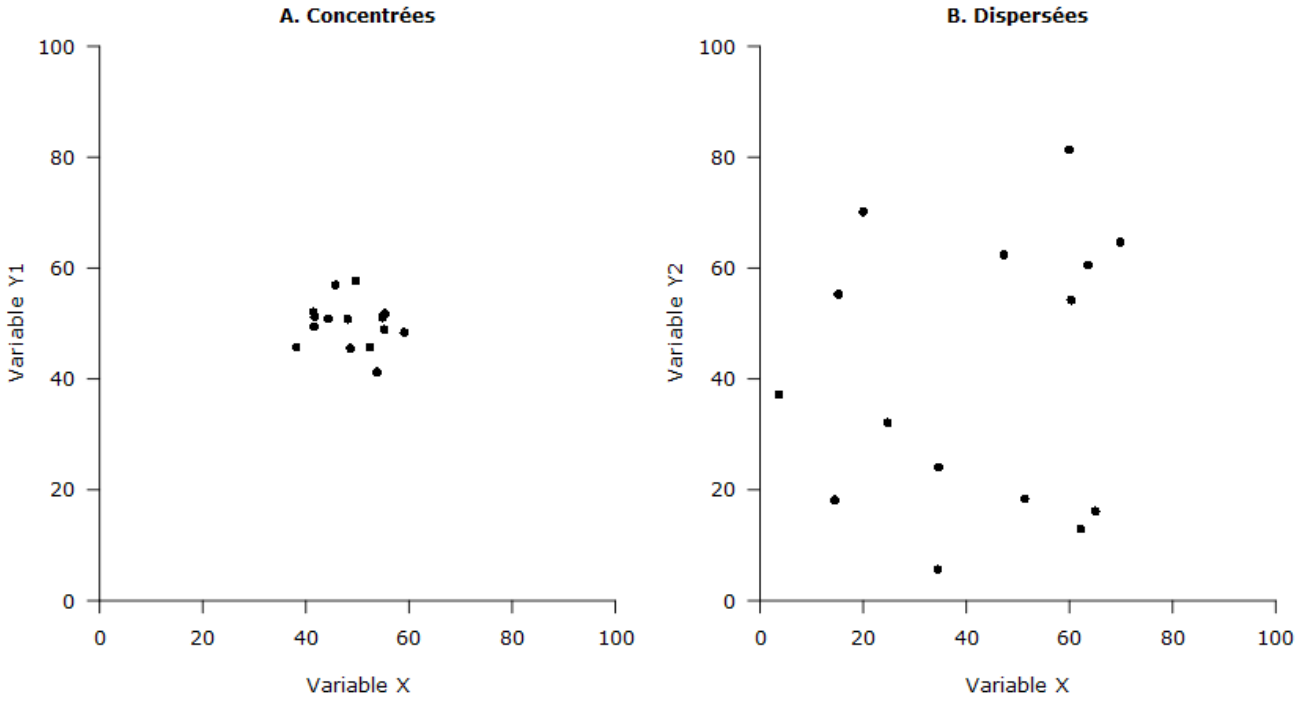
Graphique 5.6.3
Relation positive ou relation négative



Concentration ou dispersion des données

Les points peuvent être très près les uns des autres (graphique 5.6.4, partie A) ou être très dispersés dans l'espace (graphique 5.6.4, partie B).

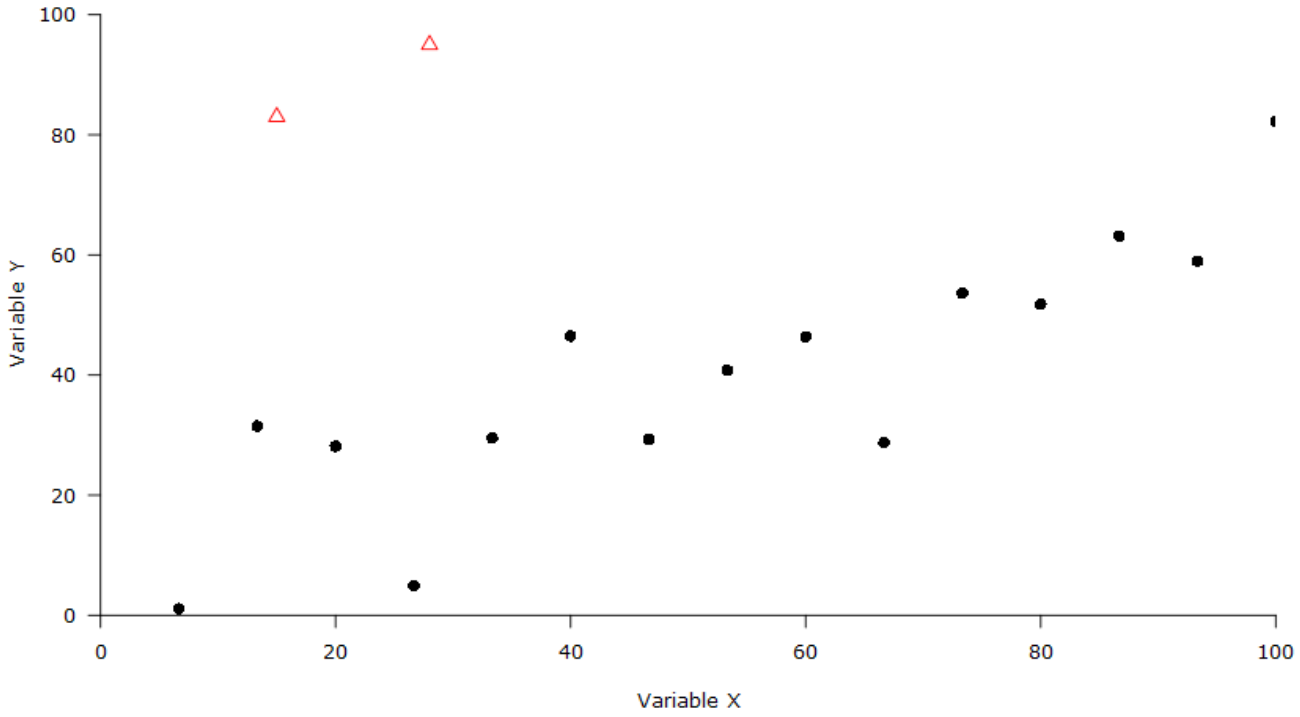
Graphique 5.6.4
Données concentrées ou données dispersées



Présence de valeurs extrêmes

En plus de montrer la relation entre deux variables, un nuage de points peut également montrer si des valeurs extrêmes sont présentes dans l'ensemble de données. Les valeurs extrêmes sont celles qui sont éloignées des autres données de l'ensemble de données, comme les deux points en rouge au graphique 5.6.5.

Graphique 5.6.5
Présence de données extrêmes

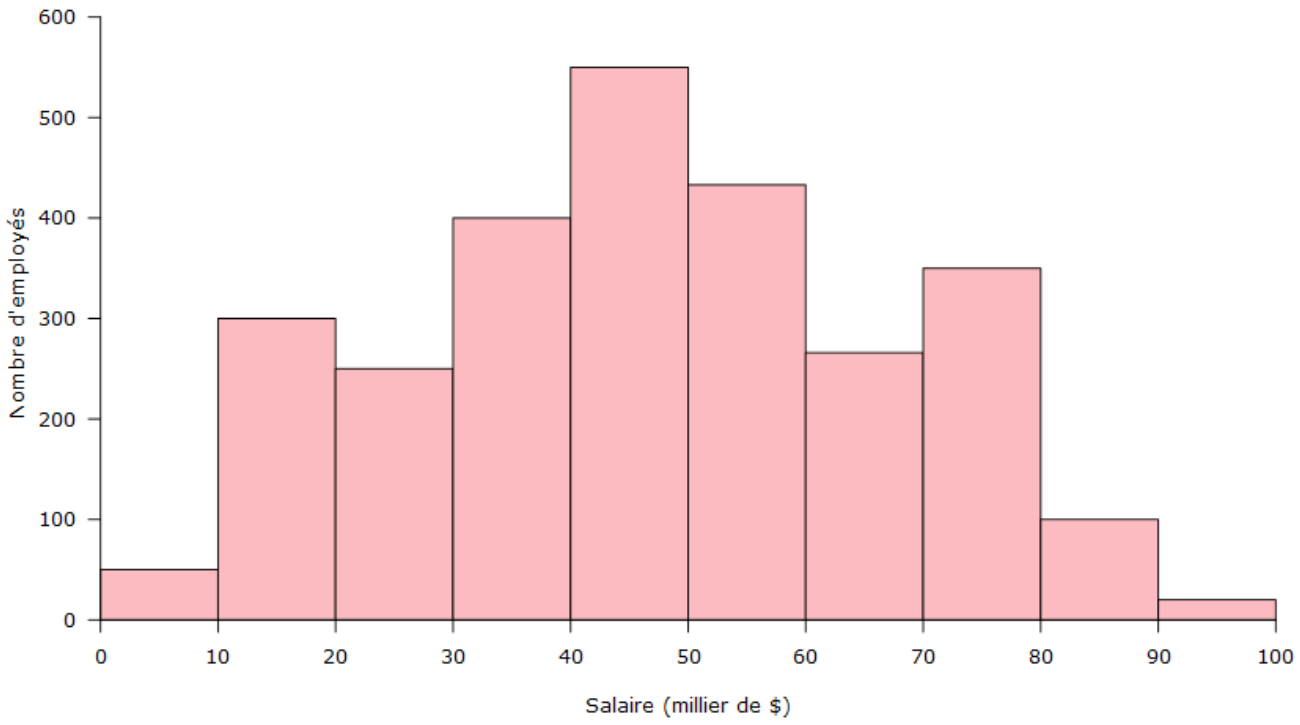


5.7 Histogramme

L'histogramme est un outil fréquemment utilisé pour résumer des données discrètes ou continues qui sont présentées par intervalles de valeurs. Il est souvent employé pour montrer les caractéristiques principales de la distribution des données de façon pratique. Il est utile pour résumer de grands ensembles de données (plus de 100 observations). Il peut également faciliter la détection d'observations inhabituelles (valeurs aberrantes) ou les intervalles sans point de donnée.

Un histogramme sépare les valeurs possibles des données en classes ou groupes. Pour chaque groupe, on construit un rectangle dont la base correspond aux valeurs de ce groupe et la hauteur correspond au nombre d'observations dans le groupe. L'histogramme a une apparence semblable au graphique à barres verticales, mais il n'y a pas d'écart entre les barres. En règle générale, l'histogramme possède des barres d'une largeur égale. Le graphique 5.7.1 est un exemple d'histogramme qui montre la distribution du revenu, une variable continue, parmi les employés d'une compagnie.

Graphique 5.7.1
Distribution des salaires des employés de la societe ABC



Le tableau suivant présente les différences entre un histogramme et un graphique à barres verticales.

Table 5.7.1
Différence entre le graphique à barres et l'histogramme

Termes de comparaison	Graphique à barres	Histogramme
Utilisation	Pour comparer différentes catégories.	Pour afficher la distribution d'une variable.
Type de variable	Variables catégoriques	Variables numériques
Apparence	La fréquence de chaque catégorie est illustrée par une barre distincte.	L'étendue des valeurs est divisée en une série d'intervalles qui ne se chevauchent pas. Les points de données sont regroupés et le nombre de points dans chaque intervalle correspond à une barre distincte.
Espace entre les barres	Il peut y avoir de l'espace entre les barres.	Il n'y a pas d'espace entre les barres.
Réorganisation des barres	L'ordre peut être modifié pour les variables nominales.	Impossible de modifier l'ordre des intervalles.

5.8 Exercices

1. Voici le nombre de parties de basketball auxquelles ont assisté 50 abonnés :

15, 10, 17, 11, 15, 12, 13, 16, 12, 14, 14, 16, 15, 18, 11, 16, 13, 17, 12, 16, 18, 15, 17, 15, 19, 13, 14, 17, 16, 15, 12, 11, 17, 16, 15, 10, 14, 15, 13, 16, 18, 15, 17, 11, 14, 17, 15, 14, 13, 16.

a. Comptez les données et présentez-les dans un tableau de distribution de fréquences.

b. Dessinez un graphique à barres verticales.

c. Décrivez les données en utilisant le résumé en cinq nombres, l'étendue et l'écart interquartile. Ces concepts ont été décrits dans la section 4 portant sur l'exploration des données.

5.9 Réponses

1. a.

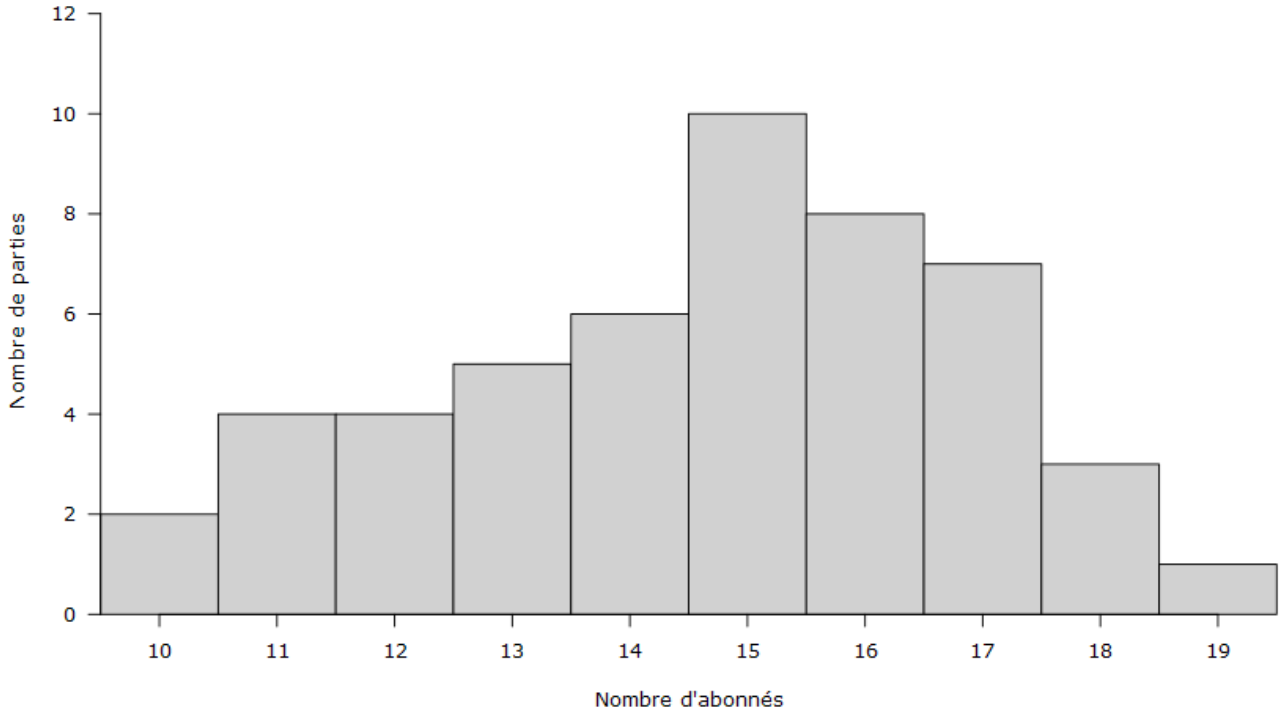
Table 5.9.1
Nombre de parties de basketball auxquelles ont assisté 50 abonnés

Nombre de parties (x)	Fréquence (f)	Pourcentage (%)	Fréquence cumulée	Pourcentage cumulé (%)
10	2	4	2	4
11	4	8	6	12
12	4	8	10	20
13	5	10	15	30
14	6	12	21	42
15	10	20	31	62
16	8	16	39	78
17	7	14	46	92
18	3	6	49	98
19	1	2	50	100
TOTAL	50	100	…	…

… n'ayant pas lieu de figurer

b.

Graphique 5.9.1
Nombre de parties de basketball auxquelles ont assisté 50 abonnés



c. Le résumé en cinq nombres est

- Minimum : 10
- Quartile inférieur : 13
- Médiane : 15
- Quartile supérieur : 16
- Maximum : 19

L'étendue est de 9 et l'écart interquartile de 3. Plus de 50 % des données se trouvent dans l'intervalle 13 à 16. La valeur la plus fréquente (mode) est 15.

Bibliographie

Australian Bureau of Statistics (1998). *Statistics – A Powerful Edge!* (2e éd). https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1331.01996?OpenDocument (site consulté le 8 décembre 2021).

Beaumont, J.-F. (2020). Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ? *Techniques d'enquête*, Statistique Canada, n° 12-001-X au catalogue, vol.46, n° 1. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2020001/article/00001-fra.htm

Lorh, S.L. (2019). *Sampling: Design and Analysis* (2e éd). Chapman & Hall/CRC Press.

Statistique Canada (2003). *Méthodes et pratiques d'enquête*. Catalogue n° 12-587-XPF. https://www150.statcan.gc.ca/n1/fr/catalogue/12-587-X

Statistique Canada (2019). *Foire aux questions sur l'utilisation de nouvelles données et de données existantes pour produire des statistiques officielles*. https://www.statcan.gc.ca/fr/nos-donnees/faq (site consulté le 8 décembre 2021).

Statistique Canada (2019). *Statistique Canada : lignes directrices concernant la qualité*. Catalogue n° 12-539-X. https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-fra.htm

Statistique Canada (2020). *Qualité des données en six dimensions* [Vidéo]. Catalogue n° 892000062020001. https://www.statcan.gc.ca/fr/afc/litteratie-donnees/catalogue/892000062020001

Statistique Canada (2020). *Que sont les données? Introduction à la terminologie et aux concepts relatifs aux données* [Vidéo]. Catalogue n° 892000062020006. https://www.statcan.gc.ca/fr/afc/litteratie-donnees/catalogue/892000062020006

Statistique Canada (2020). *Types de données : comprendre et explorer les données* [Vidéo]. Catalogue n° 892000062020004. https://www.statcan.gc.ca/fr/afc/litteratie-donnees/catalogue/892000062020004

Statistique Canada (2021). *Statistique 101 : Corrélation et causalité* [Vidéo]. Catalogue n° 892000062021002. https://www.statcan.gc.ca/fr/afc/litteratie-donnees/catalogue/892000062021002

Statistique Canada (2021). *Statistique 101 : Explorer les mesures de la dispersion* [Vidéo]. Catalogue n° 892000062020003. https://www.statcan.gc.ca/fr/afc/litteratie-donnees/catalogue/892000062020003

Statistique Canada (2021). *Statistique 101 : Explorer les mesures de la tendance centrale* [Vidéo]. Catalogue n° 892000062020002. https://www.statcan.gc.ca/fra/afc/litteratie-donnees/catalogue/892000062020002

Statistique Canada (2021). *Statistique 101 : Proportions, ratios et taux* [Vidéo]. Catalogue n° 892000062021003. https://www.statcan.gc.ca/fr/afc/litteratie-donnees/catalogue/892000062021003

Glossaire

Les définitions qui suivent visent à renseigner ceux qui ont des questions concernant certains termes utilisés en statistique, mais qui n'ont pas besoin d'une définition hautement technique. Ces définitions représentent parfois une grande simplification de notions très complexes. Pour obtenir des explications plus détaillées, vous pouvez consulter les références fournies sur la page Bibliographie.

A

Approche participative

Approche qui consiste à recueillir des renseignements provenant d'une vaste communauté d'utilisateurs et qui repose sur le principe selon lequel chaque citoyen est un expert dans son milieu.

B

Base de données

Ensemble structuré d'éléments d'information, généralement sous forme de tables.

Boîte à moustaches

Type de diagramme qui permet de visualiser le résumé en cinq nombres, soit le minimum, le quartile inférieur, la médiane, le quartile supérieur et le maximum. **Synonymes : diagramme en boîte, diagramme de quartiles**.

C

Codage de données

Processus qui vise à assigner une valeur (un code) à une réponse. Le code peut être une valeur numérique ou une chaîne de caractère.

Coefficient de variation

Rapport entre l'erreur type de l'estimation et la valeur moyenne de l'estimation sur l'ensemble des échantillons possibles.

Couplage d'enregistrements

Processus par lequel des enregistrements ou des unités provenant de différentes sources de données sont réunis dans un seul fichier à l'aide d'identifiants non uniques, tels que des noms, des dates de naissance, des adresses et d'autres caractéristiques. **Synonymes : appariement des données, jumelage des données, résolution d'entités**.

D

Dispersion

Mesure de l'étalement d'une distribution de données autour de la tendance centrale.

Distribution de fréquences

Tableau ou graphique montrant combien de fois chaque valeur ou chaque intervalle de valeurs d'une variable apparaissent dans un ensemble de données.

Données

Faits, chiffres, observations ou enregistrements qui peuvent se présenter sous la forme d'image, de son, de texte ou de mesure physique (distance, poids, longueur d'onde, etc.). Les données peuvent être collectées et traitées dans le but de tirer des conclusions.

Données administratives

Données qui sont collectées par des organismes dans le cadre de leurs opérations quotidiennes.

Données agrégées

Ensemble de données dans lequel un enregistrement est un résumé de plusieurs unités d'observation.

Données non structurées

Données qui ne sont pas organisées selon un modèle prédéfini.

Données ouvertes

Données structurées, directement exploitables par un ordinateur, qui sont partagées gratuitement et qui peuvent être utilisées sans restriction.

Données structurées

Données qui sont organisées en éléments prédéfinis, chacun correspondant à un concept ou à un élément d'information spécifique.

E

Écart interquartile

Étendue du 50 % des données qui sont au centre de la distribution, c'est-à-dire la différence entre le quartile supérieur et le quartile inférieur.

Écart semi-interquartile

Moitié de l'écart interquartile.

Écart-type

Racine carrée de la variance.

Échantillon

Sous-ensemble des unités d'une population.

Élément d'information

Plus petite pièce d'information que l'on peut collecter d'une source d'information.

Enquête

N'importe quelle activité de collecte d'information organisée et méthodique à propos des caractéristiques des unités d'une population. Le mot **enquête** est souvent utilisé pour faire référence à une enquête-échantillon, par opposition à un recensement.

Enquête-échantillon

Enquête dont les données sont collectées seulement pour certaines unités d'une population cible.

Ensemble de données

Regroupement de données qui ont en commun les définitions des unités d'observation et des variables. **Synonyme : jeu de données**.

Erreur due à l'échantillonnage

Différence entre l'estimation dérivée d'une enquête par sondage et la vraie valeur qui serait obtenue si un recensement de la population entière était effectué dans les mêmes conditions.

Erreur non due à l'échantillonnage

Toutes les sources d'erreur qui ne sont pas liées à l'échantillonnage.

Erreur type

Racine carrée de la variance échantillonnale.

Étendue

Différence entre la plus petite valeur (minimum) et la plus grande valeur (maximum).

F

Feuille de calcul

Feuille de travail créée par un tableur, dans laquelle on entre des données et qui permet d'effectuer des calculs simples et complexes.

Fichier texte délimité

Fichier texte utilisé pour stocker des données, dans lequel chaque ligne représente une unité et chaque ligne présente des champs séparés par un délimiteur. Les délimiteurs les plus communs sont la virgule, le point-virgule et la tabulation.

Fréquence

Nombre de fois qu'une valeur apparaît dans un ensemble de données. Il peut également s'agir du nombre d'évènements ou d'items. **Synonymes : compte**.

Fournisseur de données

Individus ou organisations qui collectent et traitent les données parce qu'ils ont besoin d'information, et qui rendent accessibles ces données aux utilisateurs des données.

I

Imputation des données

Processus utilisé pour assigner des valeurs de remplacement aux valeurs manquantes, invalides ou incohérentes qui ont été identifiées lors de la vérification des données.

Information statistique

Données qui ont été enregistrées, classées, organisées, reliées ou interprétées à l'intérieur d'un cadre conceptuel de sorte qu'une signification en a émergé. **Synonyme : renseignement statistique**.

Intervalle de confiance

Intervalle de valeurs autour de l'estimation qui a une certaine probabilité d'inclure la vraie valeur de la mesure d'intérêt dans la population.

M

Marge d'erreur

Moitié de la largeur de l'intervalle de confiance associé à une estimation.

Médiane

Point milieu d'un jeu de données, de sorte que 50 % des unités ont une valeur inférieure ou égale à la médiane et 50 % des unités ont une valeur supérieure ou égale. **Synonyme : deuxième quartile**.

Mégadonnées

Ensemble de données dont le nombre d'enregistrements et le nombre de variables sont si élevés qu'ils dépassent les capacités des logiciels traditionnels à traiter l'information en un temps raisonnable.

Métadonnées

Données à propos des données, incluant la description des données, la propriété, les chemins d'accès, les droits d'accès, la qualité et d'autres informations pour les données mettre en contexte.

Microdonnées

Ensemble de données dans lequel un enregistrement représente une seule unité d'observation.

Mode

Pour les variables catégoriques ou discrètes, il s'agit de la valeur ou des valeurs qui correspondent à la fréquence maximale observée. Pour les variables continues, les intervalles de classe modale correspondent aux sommets de la distribution de fréquences. Lorsqu'il est unique, le mode est une mesure de tendance centrale.

Moissonnage du web

Processus par lequel des renseignements sont recueillis et copiés à partir du web aux fins d'analyses ultérieures.

Moyenne

Mesure de tendance centrale qui correspond à la somme de l'ensemble des valeurs divisée par le nombre de valeurs.

Q

Quartile inférieur

Valeur au-dessous de laquelle se trouvent 25 % des données lorsqu'elles sont arrangées en ordre croissant. **Synonyme : premier quartile**.

Quartile supérieur

Valeur au-dessous de laquelle se trouvent 75 % des données lorsqu'elles sont arrangées en ordre croissant. **Synonyme : troisième quartile**.

Question fermée

Dans un questionnaire, une question fermée propose au répondant une liste de réponses prédéfinies et le répondant doit sélectionner une ou plusieurs réponses dans la liste.

Question ouverte

Dans un questionnaire, une question ouverte donne au répondant l'occasion de répondre à la question dans ses propres mots.

Questionnaire

Série de questions conçues pour l'obtention de renseignements sur un ou plusieurs sujets auprès d'un répondant.

R

Recensement

En général, enquête qui vise à collecter des données pour toutes les unités d'une population. Les recensements sont également utilisés pour lister et dénombrer les unités d'une population.

Registre statistique

Ensembles de données créés à des fins statistiques qui sont continuellement mises à jour avec des renseignements sur toutes les unités d'une population.

Renseignement statistique

Voir **Information statistique**.

S

Saisie des données

Processus qui permet de convertir les données dans un format exploitable par un ordinateur.

Source de données primaires

Les données d'une source primaire ont été collectées dans le but de produire des statistiques et de l'information statistique.

Source de données secondaires

Les données d'une source secondaire ont été collectées dans un but autre que celui de produire de l'information statistique.

Statistiques

Type d'information obtenu en soumettant les valeurs à des opérations mathématiques.

T

Télédétection

Acquisition à distance de renseignements à propos d'un objet ou d'un phénomène.

Tendance centrale

Mesure de l'emplacement où se trouve le milieu ou le centre d'une distribution.

Traitement des données

Transformation des données brutes, de façon à pouvoir les utiliser pour produire des estimations ou différentes analyses.

V

Valeur manquante

Point de données vierge ou absent.

Variable

Caractéristique mesurable qui peut prendre différentes valeurs.

Variable catégorique

Caractéristique qui n'est pas quantifiable. **Synonyme : variable qualitative**.

Variable continue

Variable numérique qui peut prendre un nombre infini de valeurs réelles possibles à l'intérieur d'un intervalle donné.

Variable discrète

Variable numérique qui ne peut prendre qu'un nombre fini de valeurs réelles possibles à l'intérieur d'un intervalle donné. Les valeurs possibles peuvent être énumérées et comptées.

Variable nominale

Variable catégorique qui décrit un nom, une étiquette ou une catégorie sans ordre naturel.

Variable numérique

Caractéristique quantifiable dont les valeurs sont des nombres. **Synonyme : variable quantitative**.

Variable ordinale

Variable catégorique dont les valeurs sont définies par une relation d'ordre entre les catégories possibles.

Variance

Écart élevé au carré moyen entre chaque donnée et le centre de la distribution mesurée par la moyenne.

Variance échantillonnale

Écart élevé au carré moyen entre une estimation et la moyenne des estimations de l'ensemble des échantillons possibles.

Vérification des données

Application de contrôles pour détecter les entrées manquantes, invalides ou incohérentes ou pour indiquer les enregistrements de données qui sont potentiellement erronés.