Nº 12-004-X au catalogue
ISBN 978-0-660-26573-5

Les statistiques : le pouvoir des données!

Date de diffusion : le 3 avril, 2001

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.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2001

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Remerciements

La présente publication électronique est le fruit des efforts de bien des gens.

Nous aimerions remercier en particulier tous nos collègues de la Diffusion, des Communications, de la Méthodologie et des divisions spécialisées de Statistique Canada, qui ont aidé à l'examen de cette publication électronique.

Nous tenons aussi à exprimer notre vive reconnaissance au Bureau de la statistique de l'Australie, qui nous a permis d'utiliser la deuxième édition de la publication *Statistics — A Powerful Edge!* comme base d'élaboration de cette publication canadienne consultable en direct.

Joe Anne Legge
Agent de recherche et de rédaction
Statistique Canada

À propos de la ressource *Les statistiques : le pouvoir des données!*

Les statistiques : le pouvoir des données! est une ressource qui vous permettra de tirer pleinement parti des statistiques. Chaque chapitre a été conçu de manière à être complet en lui-même, ce qui vous permet de ne choisir que le sujet qui vous intéresse sans même avoir à lire les autres chapitres.

Bien qu'il s'agisse d'une ressource Web destinée aux élèves du secondaire qui suivent des cours de mathématiques et de sciences de l'information, elle servira également à d'autres élèves, aux enseignants et au grand public.

La ressource *Les statistiques : les pouvoir des données!* a été élaborée et modifiée en tenant compte des commentaires et demandes formulés par les enseignants sur les sujets appropriés et sur le temps qu'on peut y consacrer dans le contexte d'un cours.

En plus de son texte exhaustif, cette ressource Web contient des exercices qui ont pour but d'aider les élèves à mieux assimiler la matière présentée.

Cette ressource vise à aider les élèves à :

- apprendre à utiliser avec plus d'assurance l'information statistique pour mieux exécuter leurs travaux scolaires;
- reconnaître l'importance de l'information statistique dans la société d'aujourd'hui;
- apprendre à utiliser avec un esprit critique les données qui leur sont présentées.

Ces objectifs vont au cœur de la mission de Statistique Canada d'appuyer la prise de décisions éclairées des Canadiens.

La première section de cette ressource définit les concepts de bases « information, données et statistiques » tandis que la deuxième explique comment planifier un processus d'enquête dans son ensemble. Les sections suivantes couvrent chaque étape du processus, tel que l'échantillonnage, la collecte de données, la conception de diagrammes sur les résultats, etc. Enfin, on aborde l'utilisation des statistiques dans la société et les problèmes qui y sont liés, ainsi que l'importance de la confidentialité et le rôle des ordinateurs dans la production de statistiques.

La longueur et le niveau de détail d'une section ne reflète aucunement l'importance du sujet dans le processus global de l'enquête. Comme nous l'avons mentionné plus haut, cela correspond plutôt à l'importance du sujet dans le programme d'études. C'est pourquoi on trouve une section détaillée portant sur la conception de diagrammes, alors que la section sur l'estimation et la pondération est plutôt brève. Bien que la conception de diagrammes soit un outil pertinent aux élèves du secondaire, les détails sur l'estimation des données sont considérés comme étant trop techniques pour la clientèle visée.

Cette ressource sera constamment mise à jour afin que les renseignements qu'elle renferme soient aussi utiles et actuels que possible.

Dans le monde de Statistique Canada

Statistique Canada n'a pas d'énoncé écrit de mission autre que celui qui figure au début de la *Loi sur la statistique*. Néanmoins, ses employés se distinguent par leur remarquable compréhension mutuelle des valeurs de l'organisme. Ces valeurs sont les principes directeurs qui président aux activités de tous les employés et aux décisions du statisticien en chef. C'est ce que M. Fellegi, autrefois statisticien en chef, a très clairement exprimé lorsqu'il est entré en fonction en septembre 1985 :

> « Je continuerai de mettre l'accent sur les aspects essentiels du programme du Bureau (Statistique Canada), notamment : la pertinence de notre programme et de notre mandat qui est d'offrir des services; la confidentialité, la neutralité et l'excellence scientifique qui, je crois, sont absolument essentielles, puisqu'elles constituent la valeur intrinsèque de l'information que nous transmettons aux utilisateurs; le fardeau de réponse; et la coordination et l'intégration. »

La *Loi sur la statistique* nous confie expressément les tâches suivantes :

- recueillir, compiler, analyser, dépouiller et publier des renseignements statistiques sur les activités commerciales, industrielles, financières, sociales, économiques et générales de la population et sur l'état de celle-ci;
- collaborer avec les ministères à la collecte, à la compilation et à la publication de renseignements statistiques, y compris les statistiques qui découlent des activités de ces ministères;
- tenir le Recensement de la population du Canada et le Recensement de l'agriculture du Canada en conformité avec la présente loi;
- veiller à prévenir le double emploi dans la collecte des renseignements par les ministères;
- en général, favoriser et mettre au point des statistiques sociales et économiques intégrées concernant l'ensemble du Canada et chacune des provinces, et coordonner des projets pour l'intégration de telles statistiques.

Sauvegarde de l'objectivité

En sauvegardant l'objectivité, Statistique Canada contribue de façon fondamentale au fonctionnement du régime démocratique canadien. Le gouvernement et son opposition politique peuvent avoir des opinions divergentes sur les politiques, tout en s'entendant sur les données de base. Pour sa part, l'électorat peut juger du rendement d'un gouvernement par sa « feuille de route », dont rendent compte en grande partie les données établies par Statistique Canada.

Sauvegarde de la confidentialité

Dans l'information qu'il obtient auprès de milliers de particuliers et d'organismes au Canada, Statistique Canada a toujours cherché en toute priorité à sauvegarder la confidentialité des réponses individuelles des répondants. À quelques exceptions près, il est obligatoire de répondre aux questionnaires des enquêtes conformément à la *Loi sur la statistique*. Statistique Canada a également l'obligation légale de veiller à ce que les réponses individuelles demeurent tout à fait confidentielles. Aucun autre organisme public – pas même le Service canadien du renseignement de sécurité ni la Gendarmerie royale du Canada – n'a accès aux réponses individuellement identifiables.

Professionnalisme et fiabilité

Statistique Canada est fier de l'objectivité des données qu'il produit. Toutefois, pour que ses données fassent autorité, elles doivent être sûres et les utilisateurs doivent savoir qu'elles ont été établies avec le plus grand professionnalisme. Statistique Canada a tout mis en œuvre pour promouvoir la fiabilité et le professionnalisme de ses activités. De nombreux organismes à l'échelle internationale ont dit que Statistique Canada comptait parmi les meilleurs organismes statistiques au monde.

Accent sur l'analyse

Comme Statistique Canada a mis l'accent sur l'analyse au milieu des années 1970, l'importance de cette activité s'est accrue au sein de la diffusion de l'information. L'analyse joue aussi un rôle en assurant la pertinence des programmes et des priorités de l'organisme et en améliorant les communications avec les utilisateurs. Pour citer encore une fois l'ancien Statisticien en chef du Canada, M. Fellegi :

> « L'analyse aide aussi à populariser nos informations. Elle fait ressortir des détails précis dans cette mer de données que nous produisons, ce qui permet non seulement de fournir aux Canadiens et Canadiennes des renseignements qu'ils peuvent utiliser, mais aussi de leur faire connaître le travail qu'accomplit notre organisme. »

Allégement du fardeau de réponse

Statistique Canada a toujours été conscient du fardeau de réponse qu'il impose aux gens, notamment aux petites entreprises.

Le Bureau tente de contrôler le caractère intrusif de ses enquêtes au moyen des méthodes suivantes :

- réduire le volume de questions en utilisant des renseignements qui ont déjà été recueillis pour des fins administratifs, comme les données fiscales ou des douanes;
- raccourcir et simplifier les questionnaires et réduire le nombre d'enquêtes;
- faire la collecte conjointe de données, par exemple, au moyen d'une enquête conjointe fédérale-provinciale;
- établir des lignes directrices pour déterminer si une enquête devrait être volontaire ou obligatoire.

Données, information et statistiques

En ce début du nouveau millénaire, nous devons constamment relever de nouveaux défis stimulants. Plus que jamais auparavant, les administrations publiques, l'industrie et le reste de la société ont besoin de données fiables qui les aideront à prendre de meilleures décisions pour s'attaquer à ces problèmes.

La nécessité d'informer la société est l'une des raisons pour lesquelles le système d'éducation canadien insiste sur le devoir pour les élèves de réunir des données et de les présenter afin de mieux exécuter leurs travaux scolaires. Toutefois, avant de se lancer dans de telles activités, il importe qu'ils comprennent bien les termes *données*, *information* et *statistiques*.

Définitions

Données

Avant la présentation et l'interprétation de renseignements, il y a la collecte et le classement des données. Autant les arbres constituent la matière première du papier, autant,les données sont la matière brute d'où naît l'information.

On pourrait définir les données comme des faits ou des chiffres se prêtant à des conclusions.

Les termes données, information et statistiques sont souvent confondus. En fait, ce sont des choses entièrement différentes, comme le démontre la figure 1.

Figure 1. Données relatives au poids recueillies auprès de 20 élèves dans votre classe

Données	Information	Statistiques
20 kg à 25 kg	5 élèves dans la catégorie de 20 à 25 kg	Poids moyen = 22,5 kg
28 kg, 30 kg	15 élèves dans la catégorie de 26 à 30 kg	Poids médian = 28 kg

Bien que les données puissent prendre différentes formes, elles sont souvent numériques. Elles peuvent ainsi porter sur une grande diversité de sujets comme :

- les mesures quotidiennes relatives au poids de chaque élève dans votre classe;
- le nombre de films loués à tous les mois par chaque ménage dans votre voisinage;
- la température de votre ville (mesurée à toutes les heures) pendant une semaine.

Il existe d'autres formes de données, qu'il s'agisse de signaux radio, d'images numérisées ou d'enregistrements laser sur disque compact.

Statistique Canada recueille des données chaque jour pour renseigner les gens.

Une fois recueillies et traitées, les données peuvent être organisées sous forme d'information. En fait, il est difficile d'imaginer pourquoi on voudrait recueillir des données à des fins autres que pour fournir de l'information. L'information engendre la connaissance de questions d'intérêt et aide les personnes et les groupes à prendre des décisions éclairées.

Dans la pratique, des décisions éclairées peuvent faire épargner des millions de dollars à un pays, si celui-ci peut, par exemple, cibler avec précision les dépenses de l'état. Grâce à l'information, il peut également y avoir des percées en médecine qui sauveront des vies, tout comme de meilleures mesures de conservation des milieux naturels de la planète.

Information

L'information pourrait être définie comme des données qui ont été consignées, classées, organisées, raccordées ou interprétées dans un cadre qui en dégage le sens.

L'information, comme les données, peut prendre différentes formes. Voici quelques exemples du genre d'information qu'on peut tirer des données.

- le nombre de personnes dans un groupe selon la catégorie de poids;
- le nombre total de ménages qui n'ont pas loué de films au cours du dernier mois;
- le nombre de jours pendant la semaine où la température était supérieure à 20 degrés Celsius.

Certains des objets culturels les plus anciens de notre pays nous ont livré une information encore utile aujourd'hui. L'astrolabe est un ancien instrument scientifique qui date de l'an 170 avant J.-C. On s'est servi d'astrolabes planétaires complexes pour mesurer la hauteur des planètes et des étoiles et en suivre les mouvements. Par ces calculs, il était également possible d'établir la latitude et le temps.

L'astrolabe

Au XVIe siècle, on a mis au point un astrolabe plus simple à l'intention des navigateurs. L'illustration présentée ci-dessus serait l'astrolabe employé par Christophe Colomb. Celui-ci alignait l'axe horizontal de l'astrolabe sur l'horizon, puis pointait l'indicateur mobile (appelé « alidade ») vers le soleil ou l'étoile polaire pour relever la position sur son cadran extérieur. Cela permettait ensuite de rapporter ces mesures aux tables astronomiques pour l'obtention de données de latitude. Les astronomes exploitent l'information sous forme d'ondes radio structurées pour explorer l'espace. Au cours des dernières années, l'information visuelle transmise par les supports télévisuel et vidéo est devenue très courante.

Au XVIIe siècle, le philosophe anglais Francis Bacon a reconnu l'importance du savoir. La citation que nous rapportons ci-après provient de lui; elle est probablement plus pertinente aujourd'hui que jamais et donne une importante raison pour laquelle le grand public devrait avoir accès à l'information.

> « La connaissance même est source de pouvoir. »
> *—Francis Bacon*

Statistiques

Les statistiques sont utilisées couramment pour la présentation de l'information. Le terme se réfère d'ordinaire à des données numériques mais il peut aussi faire référence à la science du traitement des données numériques. Les statistiques visent avant tout à fournir des renseignements utiles à l'aide de chiffres. La statistique pourrait donc être définie comme de l'information qu'on peut obtenir à partir d'opérations mathématiques sur des données numériques.

Résultat d'une partie de hockey

Maple Leafs de Toronto	Oilers d'Edmonton
6	5

Voici les statistiques qu'on peut obtenir à partir des exemples précédents :

Statistiques obtenus

Information	Statistiques
le nombre de personnes dans un groupe selon la catégorie de poids	le poids moyen des élèves de votre classe
le nombre total de ménages qui n'ont pas loué de films au cours du dernier mois	le nombre de locations de films que votre ménage a dû faire pour se classer parmi les 5 % des ménages qui ont loué le plus grand nombre de films le mois dernier
le nombre de jours pendant la semaine où la température était supérieure à 20 degrés Celsius	les températures minimales et maximales observées chaque jour de la semaine

L'un des principaux rôles de Statistique Canada est de fournir à la collectivité canadienne des statistiques qui l'aideront à prendre des décisions éclairées. Les gouvernements, les entrepreneurs, les médecins, les agriculteurs, les enseignants et les élèves utilisent grandement l'information statistique produite par Statistique Canada.

La diffusion d'information statistique qui se distingue par son exactitude et qui fait autorité vient renforcer les sociétés modernes. Les statistiques peuvent servir de base à la prise de décisions dans divers domaines, par exemple, pour décider de l'emplacement approprié des écoles ou des hôpitaux, pour savoir combien d'argent dépenser dans les programmes de bien-être social ou même quels joueurs remplacer à la mi-temps dans une partie de football! Voici un exemple d'information statistique qui pourrait donner lieu à une prise de décisions par des instances gouvernementales :

> *« Les gains moyens des jeunes âgés de 15 à 24 ans sont à la baisse depuis 1980. Les changements observés dans l'activité économique globale ont entraîné des effets différents chez les jeunes par rapport aux autres groupes d'âge. Par exemple, durant les périodes de prospérité économique, les possibilités d'emploi pour les jeunes sont à la hausse, mais leurs gains ne semblent pas augmenter au même rythme que ceux des travailleurs plus expérimentés. À titre d'exemple, au cours des périodes de 1970 à 1980 et de 1985 à 1990, les gains moyens réels des jeunes ont augmenté, mais à un taux légèrement moins élevé que le taux global de croissance.*
>
> *Au cours des périodes difficiles de 1980 à 1985 et du début des années 1990, leurs gains moyens ont donc chuté d'environ 20 %, ce qui est beaucoup plus important que le taux global de diminution.*
>
> *En 1970 et 1980, les gains moyens des jeunes âgés de 15 à 24 ans atteignaient environ la moitié de la moyenne nationale globale. En 1995, ces gains avaient diminué à 31 % de la moyenne nationale. À la suite de ces changements, la situation des jeunes bénéficiaires d'un revenu d'emploi par rapport aux autres groupes d'âge s'est détériorée de façon constante entre 1980 et 1995.*
>
> *Les gains moyens des jeunes âgés de 15 à 24 ans en 1980 s'établissaient à 13 191 $, comparativement à 8 199 $ en 1995 pour le même groupe d'âge. »*
>
> ***Source :*** *Le Quotidien de Statistique Canada, mardi 12 mai 1998*

Exemples d'information statistique

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!

Certaines professions au Canada : Recensement 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**

Pyramides des âges

Voici un exemple de diagramme sous forme de pyramide des âges conçu à partir des estimations annuelles de la population de 2001. Ce diagramme présente des données statistiques sur la population canadienne selon le groupe d'âge et le sexe. En 2001, la population totale du Canada était de 30 007 008.

Les pyramides des âges permettent habituellement de livrer de l'information statistique sur la composition d'une population. Le diagramme montre clairement l'évolution de la génération vieillissante du baby-boom.

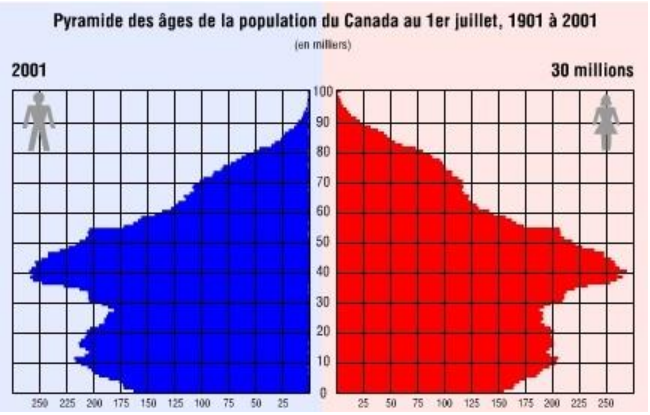


Tableau des données utilisées pour la pyramide d'âge

Pyramide animée de la population du Canada par âge

Pyramides de la population pour chaque province et territoire

Ligue canadienne de football (LCF)

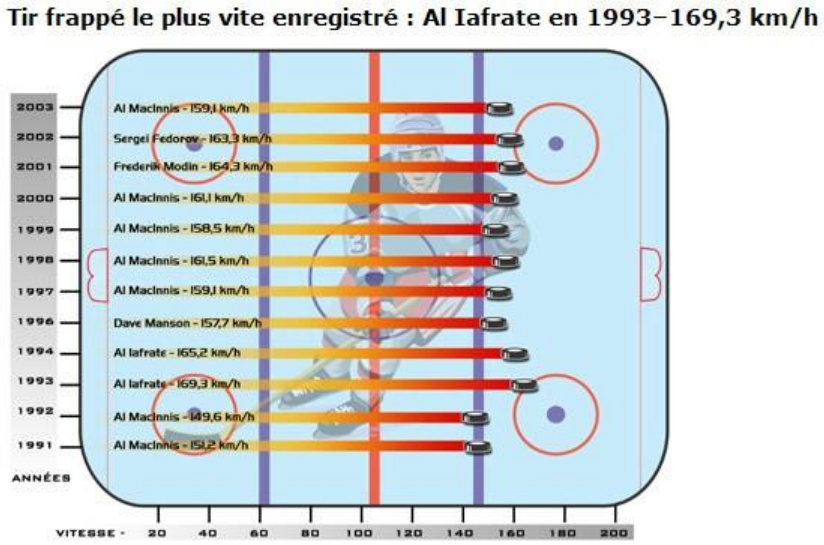
Ce tableau d'information statistique montre comment les équipes de football au Canada ont performé durant la saison de 2002 au chapitre des « verges gagnées ».

Statistiques officielles de 2002 de la Ligue canadienne de football (LCF) – total des verges gagnées

Équipes	Verges gagnées	Verges par la courses	Verges par la passe
Alouettes de Montréal	7 170	2 342	5 242
Blue Bombers de Winnipeg	6 993	1 850	5 544
Stampeders de Calgary	6 291	1 930	4 766
Eskimos d'Edmonton	6 237	2 061	4 648
Lions de la C.-B.	6 230	2 183	4 357
Tiger-Cats de Hamilton	6 139	1 558	4 807
Renegades d'Ottawa	5 701	2 518	3 600
Roughriders de la Saskatchewan	5 100	1 488	3 933
Argonauts de Toronto	4 436	1 664	3 162

Ligue nationale de hockey (LNH)

Cette illustration présente de l'information statistique sur les joueurs qui, depuis 1991, ont enregistré les tirs frappés les plus rapides au hockey. On y apprend la vitesse des tirs frappés de chaque joueur.



Plus tard, vous verrez comment il est facile de se tromper lorsqu'on interprète mal l'information statistique. Si on les utilise sagement, les statistiques peuvent être un outil puissant pour la prise de décision.

Date de modification : 2013-07-23

Exercice

1. Définissez en vos propres termes les mots *données, information* et *statistiques*. Donnez des exemples de chacun.

2. Mettez les termes suivants dans le bon ordre logique : connaissances, données, information.

3. Indiquez trois questions politiques, économiques ou sociales d'actualité sur lesquelles il est nécessaire d'obtenir de l'information. Décrivez ensuite l'information requise pour chaque question.

4. De tous les exemples d'information statistique présentés dans ce chapitre, lequel renferme le moins d'observations et lequel en contient le plus?

5. Dans l'exemple d'information sur les professions au Recensement de 1911, il n'y a aucune observation relative aux femmes pour trois professions (0 observation). Est-ce dire qu'on ne dispose pas de renseignements sur les femmes qui auraient pu occuper ces professions?

6. À l'aide de l'exemple d'information sur la Ligue canadienne de football (LCF), indiquez quelle équipe a enregistré le moins de verges gagnées au sol par la passe au cours de la saison 2002.

7. L'exemple d'information sur la Ligue nationale de hockey indique des données sur combien d'unités (joueurs de hockey)?

8. Qui, à votre avis, aurait besoin de l'information fournie à l'exemple portant sur les professions au Recensement de 1911?

9. Qui, à votre avis, aurait besoin de l'information fournie à l'exemple relatif aux pyramides de population, et dans quel but?

10. L'information présentée dans l'exemple portant sur la LCF vous dit-elle précisément quelle équipe a eu le meilleur rendement pour le total des verges gagnées au sol durant la saison 2002? Expliquez.

11. Dans quel exemple a-t-on eu besoin d'un instrument scientifique pour recueillir les données?

12. Dans quel exemple trouve-t-on toutes les observations individuelles recueillies?

Activité en classe

1. Les exemples présentés dans ce chapitre illustrent aussi la diversité des moyens de présentation de l'information statistique. On peut trouver dans les journaux ou les revues d'autres modes de présentation des statistiques. Il faut bien distinguer les données, l'information et les statistiques comme à l'exercice 1.

2. ARCHIVÉ - Recenser les Canadiens

Réponses

1.

2. L'ordre logique exact est le suivant : données, information, connaissances.

3.

4. L'exemple dans lequel on trouve le moins d'observations est celui de la Ligue canadienne de football, qui en compte huit. L'exemple des pyramides de population est celui qui compte le plus d'observations.

5. Les trois professions pour lesquelles on ne trouve aucune observation sur les femmes sont les entrepreneurs de pompes funèbres, les cochers et palefreniers et les matelots et marins. Il faut noter toutefois que la valeur 0 peut aussi être considérée comme de l'information.

6. Dans la Ligue canadienne de football, ce sont les Toronto Argonauts qui ont accumulé le moins de verges gagnées au sol par la passe (4 436) durant la saison 2002.

7. L'exemple portant sur la LNH permet d'obtenir de l'information sur 12 joueurs.

8. L'information présentée dans l'exemple portant sur les professions au Recensement de 1911 pourrait intéresser à la fois les historiens et les élèves en histoire.

9. L'information présentée dans l'exemple des pyramides de population pourrait aussi intéresser les administrations publiques pour les besoins de planification des politiques en santé et des politiques sociales.

10. Les statistiques sur le nombre total de verges gagnées ne sont pas nécessairement le reflet fidèle de la contribution globale d'une équipe.

11. Dans l'exemple portant sur la LNH, les juges ont eu recours à un vélocimètre de tirs pour capter la vitesse des tirs lancés.

12. Aucun des exemples présentés dans ce chapitre ne montre toutes les observations individuelles recueillies. (prenons l'exemple portant sur la LNH : chaque joueur a fait plus d'un tir frappé, mais on n'indique que le tir frappé le plus rapide).

Planifier une enquête

À première vue peut-être, le déroulement d'une enquête consisterait simplement 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, on devrait se poser les questions suivantes :

- 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, on doit prendre une foule de décisions sur les sujets suivants :

- 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 données

Objectifs de l'enquête

On commence par établir les objectifs qui décrivent pourquoi et pour qui on fait l'enquête. Les objectifs de l'enquête en disent long sur les données qu'on aura besoin de recueillir. Elles aident aussi à cibler la population qu'on interrogera.

> Par exemple, imaginons que le conseil étudiant de l'École secondaire St-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 définir des objectifs plus raffinés. 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 ils y trouveront ce qu'ils espèrent y trouver.)
> - Obtenir des données utiles qui serviront au comité organisateur du bal des finissants.

Le plan d'une enquête démontre comment les objectifs seront atteints : il décrit clairement la population cible, les besoins en donnée et les variables qui seront mesurées, et prévoit les questions et réponses possibles, ainsi que le traitement et l'analyse des données obtenues.

Population cible

Si l'objectif d'une enquête est d'obtenir de l'information auprès des élèves, par exemple, il faudra se demandera « 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 graduer cette année, donc, ceux qui fréquentent la dernière année du secondaire (secondaire V ou 12e année). Si certains d'entre eux étudient à temps partiel et ne prévoient pas terminer cette année, on n'aura pas besoin de les consulter. Donc la population cible serait définie comme « les élèves de l'École secondaire St-Sauveur qui fréquentent le secondaire V à temps plein ».

Il arrive quelquefois que la population cible (la population faisant l'objet de l'information à recueillir) et la population visée par l'enquête (celle qui répond à l'enquête) sont souvent différentes pour des raisons pratiques, et ce, même si elles devraient, en réalité, être identiques.

> Dans notre exemple, il se peut que certains élèves de secondaire V 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 qu'on interrogera, 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 est 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 écoles de théologie ou les collèges militaires. Ces types de diplômés seraient presque certains d'obtenir un emploi dans leur domaine respectif. 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'interviews personnelles 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 « 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 qu'on vend)
- Advenant qu'on demande 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?
- Question de sécurité, 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.)

Lorsqu'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.

En 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 volume 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 (p. ex. 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 d'exactitude de l'information que l'on vise ainsi que les ressources dont on dispose influencent le choix que l'on fait parmi les trois principaux types de collecte de données.

1. **Le recensement** est une enquête qui recueille de l'information de toutes les personnes dans un groupe ou une population.
2. **L'enquête-échantillon** recueille de l'information d'une partie seulement (d'un échantillon) d'une population. À partir des données de l'échantillon, on peut estimer les résultats pour la population dans son ensemble.
3. **Les données administratives** sont recueillies par un organisme et servent à titre de substitut d'une enquête.

Chacun présente à la fois des avantages et des inconvénients. Le choix du type de collecte dépendra de divers facteurs. Voir Types de collecte 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.

On détermine souvent le type de collecte selon le **budget** alloué pour le mener. Les coûts représentent une des principales raisons de mener une enquête-échantillon au lieu d'un recensement. Grâce à l'échantillonnage, on peut 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 aux enquêteurs de produire des données peu de temps après qu'ils en ont établi le besoin, selon un **échéancier** rapide. Par exemple, si un organisme veut mesurer le degré de sensibilisation du public suite à 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, on peut diminuer la durée de la collecte et du traitement des données tout en allouant plus de temps pour la planification de l'enquête.

Minimiser l'erreur

Lorsqu'on planifie une enquête, 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 un certain biais 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 qui en compte 25, à partir d'un échantillon de 5. 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 faisant partie de l'échantillon, moins il y aura d'erreur due à l'échantillonnage. Aussi, le fait de choisir l'échantillon de façon aléatoire tend à réduire la variation. Il est possible d'estimer l'erreur d'échantillonnage associée à différentes méthodes d'échantillonnage pour ensuite tenter de la réduire. Voir Erreur d'échantillonnage.

En choisissant de faire un recensement, on évite l'erreur liée à la variation dans l'échantillon. Mais toute enquête risque aussi d'avoir des erreurs qu'on désigne « non dues à l'échantillonnage ». Par exemple, on pourrait poser une question de façon à encourager une certaine réponse ou bien on pourrait faire une erreur

en traitant les données ou en calculant un pourcentage pour un tableau de données. On peut éviter le plus possible ce type d'erreurs en portant attention au contrôle de qualité durant toutes les étapes de l'enquête. Voir Erreur non-due à l'échantillonnage.

Taille de l'échantillon

Puisque les enquêtes-échantillons sont toutes différentes, il n'existe pas de règles rigoureuses pour déterminer la taille d'un échantillon. Les facteurs décisifs sont les suivants : l'échéancier, les coûts, les contraintes opérationnelles et la précision désirée des résultats. En évaluant chacun de ces facteurs, on pourra prendre une décision plus facilement par rapport à la taille de l'échantillon. De plus, on examinera le niveau acceptable d'erreur d'échantillonnage. S'il existe des variations considérables au sein de la population, la taille de l'échantillon devra être plus grande afin d'obtenir un degré de fiabilité précis. Voir Taille de l'échantillon.

Plan d'analyse

Après avoir identifié tous les éléments (les variables) qu'on veut mesurer et préparé le plan d'échantillonnage, la prochaine étape consiste à élaborer le plan d'analyse, c.-à-d. à prévoir les résultats des tableaux de sortie.

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 on y verra facilement les croisements proposés.

> Dans notre exemple, le comité organisateur pourrait concevoir des tableaux de résultats pour chaque variable prévue, présentés sous forme de nombre et de pourcentage (p. ex. 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 de variables, tels que « Type de musique préférée selon le sexe de l'élève ».

Ces tableaux « vides » vous aideront à vérifier si les types de questions prévues permettront de répondre aux objectifs de l'enquête. Ils illustreront de façon concrète comment l'information recueillie pourra être utilisée et si elle constituera une mesure adéquate des besoins en données.

Conception du questionnaire

C'est à partir des besoins en données et du plan d'analyse qu'on développera le questionnaire. En formulant les questions, il est utile de consulter les gens qui utiliseront les données. On pourrait aussi consulter des experts en la matière ou examiner les questions d'autres enquêtes portant sur le même sujet ou sur des thèmes semblables.

On doit veiller à ce que les questions correspondent aux objectifs de l'enquête et aux besoins en données et à ce que chaque question soit pertinente. Voir Conception des questionnaires.

Méthodes de collecte de données

L'élaboration de la méthode de collecte de données est une étape importante : on doit évaluer les coûts, les ressources matérielles et le temps requis pour mener l'enquête.

Choisissez la meilleure méthode pour recueillir les données requises. Fait à retenir : la méthode choisie aura des conséquences directes sur les coûts de l'enquête et la qualité des données. Il y a plusieurs options : l'interview personnel (sur place ou téléphonique, ces deux méthodes pouvant aussi être assisté d'un ordinateur) ou le questionnaire à remplir soi-même.

L'interview personnelle, mise en œuvre par un intervieweur qualifié, peut comporter des techniques de questionnement structuré ou non structuré. Lorsque l'interview se passe au téléphone, les questions sont structurées au moyen d'un protocole d'interview formel.

Le questionnaire à remplir soi-même doit être très structuré puisque le répondant ne pourra pas obtenir l'aide d'un intervieweur. Il retournera le questionnaire par la poste ou selon un autre mode de livraison ou encore le complètera en ligne. Voir Méthodes de collecte de données.

> Dans notre exemple, le comité organisateur pourrait opter pour l'interview personnelle administrée par des interviewers qui remplissent un questionnaire électronique dans un tableur. L'intervieweur utiliserait un ordinateur portable pour entrer les réponses des élèves dans le tableur pendant l'entrevue. Si certains élèves s'inquiètent en ce qui concerne la confidentialité de leurs réponses, l'intervieweur pourrait leur donner l'option d'entrer 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 les suivantes : le codage, la saisie de données, la mise en forme des données, le processus de traitement des données erronées ou manquantes et, si nécessaire, l'élaboration de variables calculées. Bref, l'objectif de cette étape est de produire un fichier de données exempt d'erreurs. Voir Traitement 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 donc 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, la mise en forme des données, l'essai des programmes d'ordinateur, le suivi auprès des non-répondants et la vérification des données recueillies ainsi que 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 :

- organiser les données à l'aide de tableaux de distribution de fréquence ou de diagrammes à tige et à
- feuille décrire ce qui caractérise 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 en faire la diffusion au public.

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 ce que les élèves ont rapporté.

Exercices

1. Quels sont les trois premières étapes lorsqu'on planifie une enquête?

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. Quel 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 demande la question suivante : si on introduisait de nouveaux services, comment réagiraient nos clients actuels?

3. Que peut-on faire pour réduire le plus possible les sources d'erreur dans une enquête?

Activités de classe

- Choisissez un sujet d'enquête concernant votre école, votre famille ou votre voisinage. Formulez le sujet sous la forme d'une question claire à laquelle vous voulez répondre. Dressez ensuite un plan qui déterminera comment vous recueillerez et analyserez les renseignements nécessaires pour répondre à cette question.

- Pensez à un projet d'enquête que votre classe pourrait réaliser afin d'obtenir des renseignements utiles pour un comité ou un groupe de votre école ou communauté.

Réponses

1.

a. Définir les objectifs de l'enquête – Pourquoi mène-t-on cette enquête?
b. Définir la population cible – Qui fait l'objet de l'information à recueillir?
c. Définir les besoins en données – Qu'ai-je besoin de savoir?

2.

a. Utilisation de données administratives : les renseignements seraient fournis par les médecins lorsqu'ils remplissent les certificats de 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. Vous pourriez éviter complètement l'erreur d'échantillonnage en faisant un recensement, ou bien le réduire en augmentant la taille de l'échantillon.

Vous pourriez réduire les erreurs non dues à l'échantillonnage en portant attention au contrôle de la qualité tout au long du processus d'enquête.

Méthodes d'échantillonnage

On met beaucoup l'accent dans la présente publication en ligne sur la nécessité de disposer d'information exacte pour prendre des décisions éclairées (voir la section Données, information et statistiques). Les statistiques constituent un type important d'information, et les bureaux de la statistique jouent un rôle important dans la production de cette information. Pour ce faire, ces bureaux s'appuient sur des recensements, des enquêtes-échantillon et l'utilisation de dossiers administratifs. Ces trois types de collecte de données, de même que leurs avantages et leurs inconvénients, sont décrits dans la section Collecte des données.

La présente section porte sur l'échantillonnage ainsi que sur l'estimation (le processus par lequel on applique l'information recueillie à partir d'un échantillon à l'ensemble de la population).

Sélection d'un échantillon

L'échantillonnage permet aux statisticiens de tirer des conclusions au sujet d'un tout en y examinant une partie. Il nous permet d'estimer des caractéristiques d'une population en observant directement une partie de l'ensemble de la population. Les chercheurs ne s'intéressent pas à l'échantillon lui-même, mais à ce qu'il est possible d'apprendre à partir de l'enquête et à la façon dont on peut appliquer cette information à l'ensemble de la population.

L'enquête-échantillon doit être correctement définie et organisée. Si l'on pose les mauvaises questions aux mauvaises personnes, l'information que recevront les statisticiens ne leur sera pas utile lorsqu'ils l'appliqueront à l'ensemble de la population.

Voici les étapes à suivre pour sélectionner un échantillon et s'assurer qu'il atteindra ses buts dans le cadre des activités d'un bureau national de la statistique comme Statistique Canada.

Établir les objectifs de l'enquête

La première étape de la planification d'une enquête utile et efficace consiste à en préciser de façon aussi détaillée que possible les objectifs. L'enquête ne produira probablement pas de résultats utilisables sans ces objectifs. Clarifier les objectifs de l'enquête est essentiel à son succès définitif. Il faudrait à ce stade identifier les utilisateurs initiaux et définir les utilisations initiales des données.

On devrait alors évaluer les avantages et les inconvénients d'un recensement par rapport à une enquête-échantillon ou à l'utilisation de dossiers administratifs et déterminer la méthode la plus appropriée à employer. (Nous supposerons pour l'instant qu'une enquête-échantillon est le meilleur moyen de procéder afin d'obtenir l'information dont nous avons besoin. Cette hypothèse vaudra également pendant le reste des étapes de la sélection de l'échantillon, même si un grand nombre des étapes mentionnées s'appliqueront également aux autres méthodes.)

Définir la population cible

La population cible est la population totale pour laquelle on a besoin de l'information. Par exemple, si vous devez mener une enquête sur les types de voitures les plus populaires en Saskatchewan, la population cible serait alors composée de toutes les voitures de cette province. Il faut décrire les unités qui composent la population sous forme de caractéristiques les identifiant clairement. Plus précisément, les caractéristiques suivantes définissent la population cible :

- La nature des données dont on a besoin : sur 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 de temps visée par l'enquête.
- D'autres caractéristiques, comme des caractéristiques sociodémographiques (l'intérêt vis-à-vis d'un groupe d'âge particulier, par exemple) ou le type d'industrie.

Déterminer les données à recueillir

Il faut établir les exigences de l'enquête en matière de données. On doit aussi déterminer la terminologie et les définitions nécessaires relatives aux données pour s'assurer que les exigences de l'enquête sont justifiées sur le plan opérationnel.

Fixer le degré de précision

Comme mentionné dans la section sur l'erreur d'échantillonnage, il y a un degré d'incertitude associé aux estimations établies à partir d'un échantillon. Par exemple, si vous essayez d'estimer la distance moyenne entre la maison et l'école des élèves de votre classe, qui en compte 25, à partir d'un échantillon de 5 personnes, votre estimation dépendra de l'identité des 5 élèves échantillonnés. Si les 5 élèves échantillonnés vivent tous près de l'école, les résultats ne pourront représenter la classe avec exactitude. Cette variation d'un échantillon à l'autre est ce qui cause l'erreur d'échantillonnage. Toutefois, les statisticiens peuvent estimer l'erreur d'échantillonnage associée à un plan de sondage particulier et essayer de la réduire le plus possible.

Lorsqu'on conçoit une enquête, il faut établir le degré acceptable d'incertitude des estimations découlant de l'enquête. Ce degré dépend de l'utilisation finale des résultats et de l'importance du budget global de l'enquête. Plus le budget de l'enquête sera élevé, plus on disposera de ressources et, par conséquent, moins le risque d'erreur sera élevé. De plus, si le résultat final de l'enquête consistait à servir une fin particulière, le degré acceptable d'incertitude serait alors moins élevé qu'un résultat final qui consisterait simplement à chercher des tendances générales.

La taille de l'échantillon déterminera aussi le degré d'incertitude. L'accroissement de la taille de l'échantillon entraînera une diminution de l'erreur d'échantillonnage. (Si vous échantillonnez 24 des 25 élèves de votre classe, il n'y aura pas autant de variations d'un échantillon à un autre qu'il y en aurait si vous n'échantillonniez que 5 élèves sur les 25 échantillons possibles.)

Le plan d'échantillonnage

Une fois les objectifs, les lignes directrices et les définitions élaborées, le statisticien peut travailler au plan de sondage, qui comporte trois parties :

- Le plan d'échantillonnage : la façon dont on prélèvera l'échantillon.
- Les techniques d'estimation : la façon dont on appliquera les résultats établis à partir de l'échantillon à l'ensemble de la population.

- Les mesures de la précision : la façon dont on mesurera l'erreur d'échantillonnage.

Il est question des techniques d'estimation et des mesures de la précision dans une section ultérieure. Nous étudierons, pour le moment, le plan d'échantillonnage. Les étapes suivantes amènent à définir complètement le plan d'échantillonnage :

1. Déterminer ce que sera la population observée (p. ex., 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 (p. ex., un échantillon de 100 pour une population de 1 000).
5. Sélectionner une méthode d'échantillonnage.

La population observée

On doit définir la population cible tôt durant le processus de conception de l'enquête. Il s'agit de la population pour laquelle on a besoin d'information. Il faut toutefois en exclure certains membres en raison de contraintes opérationnelles : le coût élevé de la collecte des données dans certaines régions éloignées, la difficulté d'identifier des composantes de la population cible et de les contacter, etc. Étant donné qu'il serait trop difficile, par exemple, de localiser et d'étudier chaque voiture appartenant à chacun des résidents de la Saskatchewan, on pourrait plutôt mener uniquement une enquête sur la population des principales villes et localités de cette province. Lorsque certains des membres de la population cible sont exclus d'une enquête, nous appelons la population qui est prise en compte la *population observée*. La population cible est la population que nous **voulons observer**, tandis que la population observée est la population que nous **pouvons observer**.

Ce processus a pour but de 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 gens en établissement. 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. Si l'on utilise 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 qu'on utilise pour avoir accès à la population. Il existe deux types de bases de sondage : les nomenclatures et les bases aléatoires. Une nomenclature est simplement une liste de noms et d'adresses qui donnent directement accès à des « unités » (comme une liste d'hôpitaux, une liste de restaurants et une liste d'étudiants d'une université). Les bases aléatoires sont des listes de régions géographiques qui donnent indirectement accès à des unités (comme les quartiers d'une localité). On appelle ce type d'accès un accès indirect, parce qu'il faut premièrement sélectionner une liste de régions géographiques, puis trouver le moyen d'avoir accès aux unités à l'intérieur de chaque région sélectionnée.

Supposez, par exemple, 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. Si l'on vous fournissait une base aléatoire, vous pourriez alors 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.

Lorsqu'il n'existe aucune base de sondage unique qui soit appropriée, on peut utiliser plusieurs bases de sondage. Nous traiterons plus loin de certaines techniques d'échantillonnage faisant appel aux deux types de bases de sondage.

Une bonne base de sondage devrait être complète et à jour; aucun membre de la population observée ne devrait en être exclu ni y être compté deux fois (y être représenté 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 sur la sélection de la population observée. Par exemple, si on utilise une liste de numéros de téléphone pour sélectionner un échantillon de ménages, tous les ménages n'ayant pas le téléphone seront alors exclus de la population observée.

Les unités d'enquête

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

- L'unité d'échantillonnage, qui fait partie de la base de sondage et qui peut donc être sélectionnée.
- L'unité déclarante, qui fournit l'information qu'exige l'enquête.
- L'unité de référence ou l'unité d'analyse – c'est-à-dire l'unité au sujet de laquelle de l'information est fournie – 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 suivant la base de sondage utilisée. C'est pourquoi on définit la population observée, la base de sondage et les unités d'enquête 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 peut 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, comme les fonds et le temps disponibles. Pour atteindre un certain degré de précision, il faudra que la taille de l'échantillon repose, entre autres choses, sur les facteurs suivants :

- La variabilité des caractéristiques qu'on sera en train d'observer. Si toutes les personnes membres 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 de ses membres étaient très différents, 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 importante, plus on a besoin d'un échantillon de plus grande taille. Cependant, une fois qu'on a atteint un certain niveau, une augmentation de la population n'a plus d'influence sur la taille de l'échantillon. La taille de

l'echantillon necessaire pour atteindre un certain degre de precision, par exemple, sera a peu pres la meme pour une population d'un million que pour une population deux fois plus importante.

Les methodes d'echantillonnage et d'estimation : Toutes les methodes d'echantillonnage et d'estimation ne sont pas aussi efficaces les unes que les autres. Vous aurez besoin d'un echantillon plus grand si votre methode d'echantillonnage n'est pas la technique la plus efficace. Toutefois, en raison de contraintes operationnelles et de la non-disponibilite d'une base de sondage suffisante, il se peut qu'on n'utilise pas toujours la technique la plus efficace.

La methode d'echantillonnage

Il existe deux types de methodes d'echantillonnage : L'echantillonnage probabiliste et l'echantillonnage non probabiliste. La difference entre les deux tient au fait que dans le cas de l'echantillonnage probabiliste chaque unite a une «chance» d'etre selectionnee et que cette chance peut etre quantifiee, ce qui n'est pas vrai pour l'echantillonnage non probabiliste; dans ce cas, chaque unite incluse a l'interieur d'une population n'a pas une chance egale d'etre selectionnee. La section suivante decrit les caracteristiques des deux types d'echantillonnages et fournit des details sur certaines des methodes reliees a chaque type.

Échantillonnage probabiliste

L'échantillonnage probabiliste entraîne la sélection d'un échantillon à partir d'une population, sélection qui repose sur le principe de la randomisation (la sélection au hasard ou aléatoire) ou la chance. Il est plus complexe, prend plus de temps et est habituellement plus coûteux que l'échantillonnagenon probabiliste. Toutefois, comme les unités de la population sont sélectionnées au hasard et qu'il est possible de calculer la probabilité d'inclusion de chaque unité dans l'échantillon, on peut, grâce à l'échantillonnage probabiliste, produire des estimations fiables, de même que des estimations de l'erreur d'échantillonnage et faire des inférences au sujet de la population.

Il existe plusieurs méthodes différentes permettant de sélectionner un échantillon probabiliste. La méthode qu'on choisira dépendra d'un certain nombre de facteurs, comme la base de sondage dont on disposera, la façon dont la population sera distribuée, ce que sonder les membres de la population coûtera et la façon également dont les utilisateurs analyseront les données. Lorsque vous choisirez un plan d'échantillonnage probabiliste, votre but devrait consister à réduire le plus possible l'erreur d'échantillonnage des estimations pour les variables d'enquête les plus importantes, tout en réduisant le plus possible également le délai et le coût de réalisation de l'enquête.

Voici les méthodes d'échantillonnage probabiliste les plus courantes :

- l'échantillonnage aléatoire simple;
- l'échantillonnage systématique;
- l'échantillonnage avec probabilité proportionnelle à la taille;
- l'échantillonnage stratifié;
- ll'échantillonnage en grappes;
- l'échantillonnage à plusieurs degrés;
- l'échantillonnage à plusieurs phases.

Échantillonnage aléatoire simple

Dans un *échantillonnage aléatoire simple* (EAS), chaque membre d'une population a une chance égale d'être inclus à l'intérieur de l'échantillon. Chaque combinaison de membres de la population a aussi une chance égale de composer l'échantillon. Ces deux propriétés sont ce qui définit un échantillonnage aléatoire simple. Vous devez dresser une liste de toutes les unités incluses dans la population observée pour sélectionner un échantillon aléatoire simple.

Exemple n° 1 : Il faudrait numéroter dans un ordre séquentiel chaque entrée ou inscription pour prélever un échantillon aléatoire simple d'un annuaire téléphonique. S'il y avait 10 000 entrées dans l'annuaire téléphonique et si la taille de l'échantillon était 2 000 numéros, un ordinateur devrait alors générer au hasard 2 000 numéros entre 1 et 10 000. Chaque numéro aurait la même chance qu'un autre d'être généré par l'ordinateur (ce qui respecterait l'exigence de l'échantillonnage aléatoire simple : une chance égale pour chaque unité). Les 2 000 entrées dans l'annuaire téléphonique correspondant aux 2 000 numéros aléatoires générés par l'ordinateur composeraient l'échantillon.

Un échantillonnage aléatoire simple peut s'effectuer avec ou sans remplacement. Un échantillon avec remplacement signifie qu'il est possible que l'entrée dans l'annuaire téléphonique échantillonnée soit sélectionnée deux fois ou plus. Habituellement, l'échantillonnage aléatoire simple est effectué sans remplacement, parce qu'il est plus pratique et donne des résultats plus précis. Nous ferons référence à l'échantillonnage sans remplacement lorsque nous traiterons de l'échantillonnage aléatoire simple aux fins des présentes descriptions.

L'échantillonnage aléatoire simple est la méthode d'échantillonnage la plus facile à appliquer et la plus couramment utilisée. L'avantage de cette technique tient au fait qu'elle n'exige pas de données additionnelles dans la base de sondage (comme des régions géographiques) autres que la liste complète des membres de la population observée et l'information pour les contacter. Également, puisque l'échantillonnage aléatoire simple 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.

D'un autre côté, cette technique ne fait aucunement appel aux données auxiliaires contenues dans la base de sondage (p.ex., le nombre d'employés de chaque entreprise) qui pourraient rendre le plan d'échantillonnage plus efficace. En outre, même s'il est facile d'appliquer l'échantillonnage aléatoire simple à de petites populations, le faire peut être coûteux et irréalisable pour de grandes populations, parce qu'il faut en identifier et en étiqueter toutes les unités avant l'échantillonnage. Son application peut également être coûteuse s'il faut effectuer des interviews sur place, puisqu'il est possible que l'échantillon soit géographiquement distribué dans toute la population.

Un tirage à la loterie est un bon exemple d'échantillonnage aléatoire simple. Par exemple, lorsqu'un échantillon de six numéros est généré au hasard à partir d'une population de 49 numéros, chacun de ces derniers a une chance égale d'être sélectionné et chaque combinaison de six numéros a la même chance d'être la combinaison gagnante. Même si les gens tendent à éviter une combinaison comme 1-2-3-4-5-6, cette combinaison a la même chance d'être la série gagnante de numéros que la combinaison 8-15-21-28-32-40.

Exemple n° 2 : Supposez que votre école compte 500 élèves et que vous devez mener une courte enquête sur la qualité des aliments servis à sa cafétéria. Vous déterminez qu'un échantillon de 10 élèves devrait suffire à vos fins. Pour obtenir votre échantillon, vous attribuez à chaque élève de votre école un numéro compris entre 1 et 500. Pour sélectionner cet échantillon, vous utilisez une table de numéros générés au hasard. Tout ce que vous avez à faire consiste à prendre un point de départ à l'intérieur de la table (un numéro de rangée et un numéro de colonne) et à examiner les numéros aléatoires qui y figurent. Dans ce cas, puisque les données ne dépassent pas trois chiffres, les numéros aléatoires devraient renfermer trois chiffres également. Ne tenez pas compte des numéros aléatoires supérieurs à 500, parce qu'ils ne correspondent à aucun des élèves de votre école. Rappelez-vous que votre échantillon est un échantillon sans remplacement et que, si un numéro se répète, vous devez le sauter et utiliser le numéro aléatoire suivant. Les 10 premiers numéros différents entre 001 et 500 composent votre échantillon.

Exemple n° 3 : Imaginez que vous êtes propriétaire d'un cinéma et que vous y organiserez un festival de films d'horreur le mois prochain. Pour déterminer quels films d'horreur vous y présenterez, vous voulez demander à des cinéphiles lesquels parmi les films que vous leur énumérerez ils préfèrent. Pour dresser

la liste des films nécessaire à votre sondage, vous décidez d'échantillonner 100 des 1 000 meilleurs films d'horreur de tous les temps. La population des films d'horreur se divise en proportions égales entre les films classiques (tournés en ou avant 1969) et les films modernes (produits en ou après 1970). L'une des façons d'obtenir un échantillon consisterait à écrire tous les titres des films sur des bouts de papier et à les placer dans une boîte et à tirer ensuite 100 titres et vous auriez alors votre échantillon. En utilisant cette méthode, vous auriez l'assurance que chaque film courrait une chance égale d'être sélectionné.

Vous pourriez aussi calculer la probabilité de sélection de tel ou tel film. Puisque nous connaissons la taille de l'échantillon (**n**) et la population totale (**N**), calculer la probabilité pour un film d'horreur d'être inclus dans votre échantillon deviendrait une simple question de division :

Probabilité de sélection (la même pour chaque film)
= (**n ÷ N**) x 100 %
= (100 ÷ 1 000) x 100 %
= 10 %

Cela signifie que chaque titre de film inscrit sur votre liste aurait 10 % de chances ou 1 chance sur 10 d'être sélectionné.

Vous pouvez constater que l'un des inconvénients de l'échantillonnage aléatoire simple (parce que ce n'est pas le seul, mais c'en est un important) est le fait que vous savez que la population se compose de 500 films classiques et de 500 films modernes et que vous connaissez la date de parution sur les écrans de chaque film à partir de la base de sondage, mais qu'on n'utilise aucunement cette information. L'échantillon en question pourrait renfermer 77 films classiques et 23 films modernes, qui ne seraient pas représentatifs de la population entière des films d'horreur.

Il existe des moyens de surmonter ce problème de non-représentativité (moyens dont nous traiterons brièvement dans la section Estimation). Cependant, il y a aussi des moyens de tenir compte de cette information (moyens dont nous traiterons aussi plus loin, dans la section Échantillonnagestratifié).

Échantillonnage systématique

Parfois appelé *échantillonnage par intervalles*, l'*échantillonnage systématique* (SYS) signifie qu'il existe un écart, ou un intervalle, entre chaque unité sélectionnée qui est incluse dans l'échantillon. Vous devez suivre les étapes énumérées ci-dessous pour sélectionner un échantillon systématique :

1. Numéroter de 1 à **N** les unités incluses dans votre base de sondage (où **N** est la taille de la population totale).

2. Déterminer l'intervalle d'échantillonnage (**K**) en divisant le nombre d'unités incluses dans la population 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. K = 4, par conséquent. Vous devrez sélectionner une unité sur 4 pour avoir finalement au total 100 unités à l'intérieur de votre échantillon.

3. Sélectionner au hasard un nombre entre 1 et **K**. Ce nombre s'appelle l'origine choisie au hasard et serait le premier nombre inclus dans votre échantillon. À l'aide de l'échantillon fourni ci-dessus, vous sélectionneriez un chiffre entre 1 et 4 à partir d'une table de nombres aléatoires (pris au hasard). 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. Sélectionner chaque **K^{e}** (dans ce cas, 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, à l'aide de l'exemple fourni ci-dessus, que dans le cas d'un échantillonnage systématique, seuls quatre échantillons possibles, qui correspondent aux quatre origines choisies au hasard également possibles, peuvent être sélectionnés :

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 membre de la population ne fait partie que de l'un des quatre échantillons et chaque échantillon a une chance égale d'être sélectionné. Cela nous permet de constater que chaque unité a une chance sur quatre d'être sélectionnée à l'intérieur de l'échantillon. Sa probabilité d'être sélectionnée est la même que si l'on sélectionnait un échantillon aléatoire simple de 100 unités. La principale différence tient au fait que dans le cas d'un échantillonnage aléatoire simple, toute 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. Cela nous permet aussi de constater à quel point l'échantillonnage systématique est précis comparativement à l'échantillonnage aléatoire simple. L'ordre de la population incluse dans la base de sondage déterminera les échantillons possibles pour l'échantillonnage systématique. Si la population est distribuée au hasard dans la base de sondage, un échantillonnage systématique devrait alors produire des résultats similaires à ceux d'un échantillonnage aléatoire simple.

On utilise souvent cette méthode dans l'industrie, où l'on sélectionne une unité pour des essais à partir d'une chaîne de production afin de s'assurer que la machinerie et l'équipement sont d'une qualité uniforme. Un essayeur à l'intérieur d'une usine pourrait, par exemple, soumettre à un contrôle de la qualité chaque 20e produit sur une ligne de montage. L'essayeur pourrait choisir une origine au hasard entre les nombres 1 et 20. Cela déterminerait le premier produit à essayer; chaque 20e produit serait ensuite soumis à des essais.

Les intervieweurs peuvent utiliser 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 magasin, après avoir sélectionné au hasard la première personne. Un enquêteur peut interviewer les occupants de chaque 5e maison d'une rue, après avoir sélectionné au hasard l'une des cinq premières maisons.

Exemple n° 4 : Imaginez que vous devez mener une enquête pour votre collège ou votre université sur les logements pour les étudiants. Dix mille (10 000) étudiants sont inscrits dans votre établissement d'enseignement et vous voulez en prélever un échantillon systématique de 500. Pour ce faire, vous devez premièrement déterminer ce que serait votre intervalle d'échantillonnage (**K**) :

Population totale ÷ taille de l'échantillon = intervalle d'échantillonnage
N ÷ n = K

= 10 000 ÷ 500
= 20

Il faudrait attribuer un numéro séquentiel à chaque étudiant pour entreprendre cet échantillonnage systématique. On choisirait le point de départ en sélectionnant un numéro au hasard entre 1 et 20. Si ce numéro était 9, on sélectionnerait alors le 9e étudiant inscrit sur la liste et chaque 20e étudiant par la suite. L'échantillon d'étudiants serait constitué de ceux qui correspondraient aux numéros d'étudiant 9, 29, 49, 69... 9 929, 9 949, 9 969 et 9 989.

Dans les exemples utilisés jusqu'ici, l'intervalle d'échantillonnage **K** était toujours un nombre entier, mais ce n'est pas toujours le cas. Par exemple, si vous prélever un échantillon de 30 unités d'une population qui en compte 740, votre intervalle d'échantillonnage (ou **K**) sera 24,7. Dans de tels cas, il existe quelques possibilités de faire en sorte que le nombre soit plus facile à utiliser. Vous pouvez arrondir le nombre, c'est-à-dire au nombre entier supérieur le plus rapproché (le plus près) ou au nombre entier inférieur le plus rapproché. Le fait de l'arrondir au nombre entier inférieur le plus rapproché vous assurera de sélectionner au moins le nombre d'unités que vous vouliez à l'origine (et vous pourrez ensuite supprimer certaines unités pour obtenir la taille exacte de l'échantillon de votre choix). Il existe des techniques pour adapter l'échantillonnage systématique dans les cas où **N** (la population totale) n'est pas un multiple de **n** (la taille de l'échantillon), mais donne encore un échantillon qui est exactement le même que les unités **n**. Nous ne traiterons pas ici de ces techniques.

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) et que l'échantillon est distribué dans des proportions égales à l'intérieur de la population répertoriée. Le plus gros inconvénient de la méthode d'échantillonnage systématique tient au fait que les échantillons possibles risquent de ne pas être représentatifs de la population s'il existe un certain cycle sur le plan du mode d'ordonnancement de la population inscrite sur une liste et si ce cycle 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° 5 : 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, y compris 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 réflexions 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 finalement 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 des réflexions de vos employés.

Échantillonnage avec probabilité proportionnelle à la taille

Pour l'échantillonnage probabiliste, il faut que chaque membre de la population observée ait une chance d'être inclus dans l'échantillon, mais il n'est pas nécessaire que cette chance 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'échantillonnage afin d'en accroître l'efficacité. Cela s'appelle l'échantillonnage avec *probabilité proportionnelle à la taille* (PPT). 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 accroisse l'efficacité. C'est une méthode d'échantillonnage plus complexe dont nous ne traiterons pas ici davantage.

Échantillonnage stratifié

Lorsqu'on utilise l'*échantillonnage stratifié*, on divise la population en groupes homogènes (appelés strates), qui sont mutuellementexclusifs, puis on sélectionne à partir de chaque strate des échantillons indépendants. On peut utiliser n'importe quelle des méthodes d'échantillonnage mentionnées dans la présente section (et il en existe d'autres) pour sélectionner l'échantillon à l'intérieur de chaque strate. La méthode d'échantillonnage peut varier d'une strate à une autre. Lorsqu'on utilise l'échantillonnage aléatoire simple pour sélectionner l'échantillon à l'intérieur de chaque strate, on appelle le plan d'échantillonnage un plan d'échantillonnage aléatoire simple stratifié. On peut stratifier avant l'échantillonnage une population au moyen de toute variable dont on dispose pour la totalité des unités incluses dans la base de sondage (comme l'âge, le sexe, la province de résidence, le revenu, etc.)

Pourquoi doit-on créer des strates? Pour bien des raisons, la principale étant que leur création peut rendre la stratégie d'échantillonnage plus efficace. Nous avons mentionné précédemment que vous aviez besoin d'un échantillon plus grand pour obtenir une estimation plus exacte d'une caractéristique qui varie beaucoup d'une unité à l'autre. Si chaque personne incluse dans une population, par exemple, 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 l'idée qui sous-tend le gain d'efficacité qu'on réalise grâce à la stratification. Si vous créez des strates à l'intérieur desquelles des unités auraient des caractéristiques qui seraient similaires (comme le revenu) et qui différeraient considérablement de celles d'unités incluses dans d'autres strates (comme la profession et le type de logement), 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 échantillonnage aléatoire simple de la population entière sans effectuer de stratification, il vous faudrait un échantillon plus grand que la totalité de tous les échantillons de strate afin d'obtenir pour le revenu total une estimation du même degré de précision.

L'échantillonnage stratifié nous assure d'obtenir une taille d'échantillon suffisante pour des sous-groupes de la population à laquelle nous nous intéressons. Étant donné que chaque strate devient une population indépendante lorsque vous stratifiez une population, vous devrez déterminer pour chaque strate la taille de l'échantillon.

Exemple n° 6 : Supposez que vous voulez estimer combien d'élèves des écoles secondaires ont un emploi à temps partiel, et ce, tant au niveau national qu'à celui de chaque province. 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 pour effectuer cette sélection), vous auriez finalement en moyenne un peu plus de 100 personnes de l'Île-du-Prince-Édouard, puisque cette province représente moins de la moitié de 1 % de toute la population canadienne. Cet échantillon ne serait probablement pas assez important pour le genre d'analyse détaillée auquel vous songeriez. Le fait de stratifier votre liste par province, en supposant encore une fois que vous disposeriez de cette information, puis de sélectionner une taille d'échantillon pour chacune des provinces vous permettrait de déterminer la taille d'échantillon exacte qu'il vous faudrait pour l'Île-du-Prince-Édouard. Ainsi, afin d'obtenir une bonne représentation de l'Île-du-Prince-Édouard, vous utiliseriez un échantillon plus important que celui que la méthode d'échantillonnage aléatoire simple lui attribuerait.

Exemple n° 7 : Un conseil scolaire ontarien voulait évaluer l'opinion des élèves sur la suppression de la 13e année du programme d'études secondaires. Il a décidé de sonder les élèves de l'école secondaire de l'Érablière et a utilisé une technique d'échantillonnage stratifié pour s'assurer de sélectionner un échantillon représentatif d'élèves de toutes les années d'études offertes dans cet établissement d'enseignement.

Dans ce cas, les strates étaient les cinq années d'études (les 9, 10, 11, 12 et 13e années). Le conseil scolaire a ensuite sélectionné un échantillon à l'intérieur de chaque strate. Les noms des élèves sélectionnés à l'intérieur de cet échantillon en ont été extraits à l'aide de la méthode d'échantillonnage aléatoire simple ou d'échantillonnage systématique, ce qui a donné un échantillon total de 100 élèves.

La stratification est des plus utiles lorsque les variables de stratification sont :

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

Échantillonnage en grappes

Il est parfois trop dispendieux de disséminer un échantillon dans l'ensemble de la population. Les coûts de déplacement risquent de devenir élevés lorsque les intervieweurs doivent sonder des gens d'un bout à l'autre du pays. Les statisticiens peuvent choisir la technique de l'échantillonnage en grappes pour réduire les coûts.

La technique de l'échantillonnage en grappes entraîne la division de la population en groupes ou en grappes comme son nom l'indique. Suivant cette technique, on sélectionne au hasard un certain nombre de grappes pour représenter la population totale, puis on englobe dans l'échantillon toutes les unités incluses à l'intérieur des grappes sélectionnées. On n'inclut dans l'échantillon aucune unité de grappes non sélectionnées; ces unités sont représentées par celles tirées de grappes sélectionnées. La technique en question diffère de la technique d'échantillonnagestratifié, qui entraîne la sélection d'unités de chaque groupe.

Mentionnons, entre autres exemples de grappes, les usines, les établissements d'enseignement et les régions géographiques telles que les subdivisions électorales. On utilise les grappes sélectionnées pour représenter la population.

Exemple n° 8 : Supposez que vous représentez une organisation d'athlétisme désirant déterminer quels sports pratiquent les élèves de 11e année au Canada. Il serait trop coûteux et trop long d'interroger chaque élève canadien de 11e année ou même deux ou trois élèves de chaque classe de 11e année au Canada. On sélectionne plutôt au hasard 100 écoles de tout le pays.

Ces écoles fournissent des grappes d'échantillons. On sonde ensuite chaque élève de 11e année de chacune des 100 grappes. Les élèves inclus dans ces grappes représentent, en effet, tous les élèves de 11e année au Canada.

Exemple n° 9 : Imaginez que le conseil municipal d'une petite localité veut faire enquête sur l'utilisation par ses résidents des services de soins de santé.

Le conseil demande premièrement à Statistique Canada des cartes des subdivisions électorales qui identifient et étiquettent chaque îlot. Il dresse à partir de ces cartes une liste de tous les îlots. Cette liste servira de base de sondage.

Chaque ménage de la localité en question fait partie d'un îlot et chaque îlot représente une grappe de ménages. Le conseil sélectionne au hasard un certain nombre d'îlots. Il dresse ensuite, à l'aide de la méthode d'échantillonnage aléatoire simple, une liste de tous les ménages inclus dans les îlots sélectionnés; ces ménages composent l'échantillon de l'enquête.

Comme nous l'avons mentionné, réduire les coûts est l'une des raisons d'utiliser l'échantillonnage en grappes. Ce dernier crée des « poches » d'unités échantillonnées, au lieu de disséminer l'échantillon sur tout le territoire. Le fait que l'on ne dispose pas parfois d'une liste de toutes les unités incluses dans la population (une nécessité lorsqu'on effectue un échantillonnagealéatoiresimple, un échantillonnagesystématique ou un échantillonnageavecprobabilité proportionnelle à la taille), tandis qu'une liste de toutes les grappes est disponible ou facile à dresser, constitue une autre raison d'utiliser l'échantillonnage en grappes.

Dans la plupart des cas, une perte d'efficacité comparativement à ce qu'il en est lorsqu'on utilise un échantillonnage aléatoire simple représente le principal inconvénient de l'emploi de l'échantillonnage en grappes. Il est habituellement 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 les deux exemples précédents, les élèves de la même école ont tendance à pratiquer les mêmes types de sports (suivant les installations dont dispose leur établissement d'enseignement); de la même façon, les gens âgés tendent à vivre dans des quartiers bien précis et à être de grands utilisateurs des services de santé.

L'échantillonnage en grappes ne permet pas de contrôler totalement la taille finale de l'échantillon, ce qui constitue un autre inconvénient de son utilisation. Puisque toutes les écoles ne comptent pas le même nombre d'élèves de 11e année et que les îlots ne renferment pas non plus le même nombre de ménages, et que vous devez sonder chaque élève ou ménage inclus dans votre échantillon, ce dernier peut être finalement d'une taille supérieure ou inférieure à ce que vous espériez.

Échantillonnage à plusieurs degrés

La méthode d'échantillonnage à plusieurs degrés ressemble à la méthoded'échantillonnageengrappes, sauf qu'il faut dans son cas prélever un échantillon à l'intérieur de chaque grappe sélectionnée, plutôt que d'inclure toutes les unités dans la grappe. Ce type d'échantillonnage exige au moins deux degrés. On identifie et sélectionne au premier degré de grands groupes ou de grandes grappes. Ces grappes renferment plus d'unités de la population qu'il n'en faut pour l'échantillon final.

Pour obtenir un échantillon final, on prélève au second degré des unités de la population à partir des grappes sélectionnées (à l'aide de l'une des méthodes d'échantillonnageprobabiliste possibles). Si l'on utilise plus de deux degrés, le processus de sélection d'unités de la population à l'intérieur des grappes se poursuit jusqu'à l'obtention d'un échantillon final.

Exemple n° 10 : Dans l'exemple n° 8 d'échantillonnage en grappes, on choisissait 100 écoles, puis on interviewait chaque élève de 11e année de ces écoles. Dans le cas de l'échantillonnage à plusieurs degrés, vous pourriez plutôt sélectionner davantage d'écoles, vous procurer une liste de tous les élèves de 11e année de ces écoles sélectionnées et choisir un échantillon au hasard (un échantillonaléatoiresimple, par exemple) d'élèves de chaque école. Ce serait un plan d'échantillonnage à deux degrés.

Vous pourriez aussi obtenir une liste de toutes les classes de 11e année des écoles sélectionnées, prélever un échantillon aléatoire de classes de chacune de ces écoles, vous procurer une liste de tous les élèves des classes sélectionnées et finalement choisir un échantillon au hasard d'élèves de chaque classe. Ce serait un plan d'échantillonnage à trois degrés. Le processus se complique chaque fois que nous ajoutons un degré.

Imaginez maintenant que chaque école compte en moyenne 80 élèves de 11e année. L'échantillonnage en grappes donnerait alors à votre organisation un échantillon d'environ 8 000 élèves (100 écoles x 80 élèves de 11e année). Si vous vouliez un échantillon de plus grande taille, vous pourriez sélectionner des écoles comptant davantage d'élèves et, pour obtenir un échantillon plus petit, sélectionner des écoles comptant moins d'élèves.

L'un des moyens de contrôler la taille de l'échantillon consisterait à stratifier les écoles en grande taille, en taille moyenne et en petite taille (tailles ici signifiant le nombre d'élèves de 11e année) et à sélectionner un échantillon d'écoles de chaque strate. On appelle cette méthode la méthode d'échantillonnage en grappes stratifiées.

Si vous utilisiez un plan d'échantillonnage à trois degrés, vous pourriez sélectionner un échantillon de 400 écoles, puis sélectionner deux classes de 11e année par école (en supposant qu'il y ait deux classes de 11e année ou plus dans chaque établissement d'enseignement). Finalement, vous pourriez 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 disséminé.

Vous pouvez constater à partir de cet exemple que l'échantillonnage à plusieurs degrés vous offre quand même l'avantage d'un échantillon plus concentré, ce qui réduit les coûts. Cependant, l'échantillon n'est pas aussi concentré que d'autres grappes et sa taille est quand même plus grande que celle d'un échantillon aléatoire simple. Vous n'avez pas non plus besoin de disposer d'une liste de tous les élèves membres de la population. Tout ce qu'il vous faut, c'est une liste des classes des 400 écoles et une liste également des élèves des 800 classes. Il est vrai que ce type d'échantillon exige plus d'information qu'en nécessite l'échantillonnage en grappes. L'échantillonnage à plusieurs degrés épargne cependant quand même beaucoup de temps et d'efforts, parce qu'il ne nécessite pas la création d'une liste de toutes les unités incluses dans une population.

Échantillonnage à plusieurs phases

Un *échantillonnage à plusieurs phases* entraîne la collecte de données de base auprès d'un échantillon d'unités de grande taille et ensuite, pour un sous-échantillon de ces unités, la collecte de données plus détaillées. 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 les similarités entre eux sur le plan de leur appellation. Même si l'échantillonnage à plusieurs phases suppose aussi le prélèvement de deux échantillons ou plus, dans son cas, tous les échantillons sont tirés de la même base de sondage et les unités sont structurellement les mêmes à chaque phase. Comme dans le cas de l'échantillonnage à plusieurs degrés, plus l'on utilisera de phases, plus le plan d'échantillonnage et l'estimation deviendront complexes.

L'échantillonnage à plusieurs phases est utile lorsqu'il manque à l'intérieur de la base de sondage des données auxiliaires qui pourraient servir à stratifier la population ou à rejeter à la sélection une partie de la population.

Exemple n° 11 : Supposez qu'une organisation a besoin d'information sur des éleveurs de bétail de l'Alberta, mais que dans la base de sondage sont énumérés tous les types d'exploitations agricoles : d'élevage de bétail et de production laitière, de grains, de porcs, de volailles et de fruits et de légumes. Pour compliquer les choses, la base de sondage ne fournit aucune donnée auxiliaire sur les exploitations agricoles qui y sont énumérées.

On pourrait 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? » Si elle ne comptait qu'une seule question, cette enquête devrait entraîner un faible coût par interview (surtout si elle était faite par téléphone), ce qui, par conséquent, permettrait à l'organisation de prélever un grand échantillon. Une fois le premier échantillon prélevé, on pourrait en extraire un second, plus petit, à partir des éleveurs de bétail et poser à ces derniers des questions plus détaillées. En utilisant cette méthode, l'organisation éviterait de dépenser de l'argent à sonder des unités ne faisant pas partie du champ d'observation (c'est-à-dire les producteurs agricoles autres que les éleveurs de bétail).

On pourrait utiliser l'échantillonnage à plusieurs phases lorsqu'on disposerait d'un budget insuffisant pour recueillir des données auprès de l'échantillon entier ou lorsque recueillir des données auprès de l'échantillon entier imposerait un fardeau excessif aux répondants ou même quand les coûts de collecte pour différentes questions d'une enquête différeraient énormément.

Exemple n° 12 : On pose aux participants à une enquête sur la santé des questions de fond au sujet de leur régime alimentaire et de leurs habitudes au niveau de la consommation de tabac et d'alcool et de la pratique de l'exercice physique. Cette enquête oblige, en outre, les répondants à se soumettre à certains examens médicaux, comme courir sur un tapis roulant ou faire mesurer leur tension artérielle et leur taux de cholestérol.

Remplir des questionnaires ou interviewer des participants sont des procédures relativement peu dispendieuses, mais les examens médicaux exigent la supervision et l'aide d'un professionnel ou d'une professionnelle de la santé qualifié(e), de même que l'utilisation d'un laboratoire équipé, ce qui peut dans un cas comme dans l'autre être assez coûteux. La meilleure façon de mener l'enquête susmentionnée consisterait à utiliser une méthode d'échantillonnage à deux phases. À la première phase, on soumettrait aux interviews un échantillon d'une taille appropriée. On prélèverait à partir de cet échantillon un échantillon plus petit. Ce sont les membres de ce second échantillon qui passeraient des examens médicaux.

Échantillonnage non probabiliste

La différence entre l'*échantillonnage probabiliste* et l'*échantillonnage non probabiliste* tient à une hypothèse de base au sujet de la nature de la population étudiée. Dans le cas de l'échantillonnage probabiliste, chaque unité a une chance d'être sélectionnée. Dans celui de l'échantillonnage non probabiliste, on suppose que la distribution des caractéristiques à l'intérieur de la population est égale. C'est ce qui fait que le chercheur croit que n'importe quel échantillon serait représentatif et que les résultats, par conséquent, seront exacts. Pour l'échantillonnage probabiliste, la randomisation est une caractéristique du processus de sélection, plutôt qu'une hypothèse au sujet de la structure de la population.

Dans le cas de l'échantillonnage non probabiliste, puisqu'on choisit arbitrairement des unités, il n'existe aucune façon d'estimer la probabilité pour une unité quelconque d'être incluse dans l'échantillon. Également, comme la méthode en question ne fournit aucunement l'assurance que chaque unité aura une chance d'être incluse dans l'échantillon, on ne peut estimer la variabilité de l'échantillonnage ni identifier le biais possible.

On ne peut mesurer la fiabilité d'un échantillonnage non probabiliste; la seule façon de mesurer la qualité des données en résultant consiste à comparer certains des résultats de l'enquête à l'information dont on dispose au sujet de la population. Encore une fois, rien ne fournit l'assurance que les estimations ne dépasseront pas un niveau acceptable d'erreur. Les statisticiens hésitent à utiliser les méthodes d'échantillonnage non probabiliste, parce qu'il n'existe aucun moyen de mesurer la précision des échantillons en découlant.

Malgré ces inconvénients, les méthodes d'échantillonnage non probabiliste peuvent être utiles lorsqu'on désire des commentaires descriptifs au sujet des échantillons eux-mêmes. Deuxièmement, leur utilisation prend peu de temps tout en étant plus économique et plus pratique. Il existe aussi des domaines, comme la recherche sociale appliquée, où il est impossible ou presque impossible d'effectuer un échantillonnage probabiliste. Statistique Canada utilise l'échantillonnage probabiliste pour presque toutes ses enquêtes, mais emploie l'échantillonnage non probabiliste afin de tester ses questionnaires et aux fins de certaines études préliminaires durant le stade d'élaboration d'une enquête.

L'application de la plupart des méthodes d'échantillonnage non probabiliste exige un certain effort et une certaine organisation, mais d'autres méthodes d'échantillonnage non probabiliste, comme l'échantillonnage de commodité, sont à l'occasion appliquées et n'exigent pas de plan d'action formel.

Voici les types les plus courants des méthodes en question :

- l'échantillonnagedecommoditéouàl'aveuglette;
- l'échantillonnagevolontaire; l'échantillonnageau
- jugé; l'échantillonnageparquotas.

Échantillonnage de commodité ou à l'aveuglette

On appelle parfois l'*échantillonnage de commodité* l'*échantillonnage à l'aveuglette* ou *accidentel*. Cet échantillonnage n'est pas normalement représentatif de la population cible, parce qu'on ne sélectionne des unités d'échantillonnage dans son cas que si on peut y avoir facilement et commodément accès.

Il arrive que monsieur ou madame Tout-le-monde utilise l'échantillonnage de commodité. Un critique gastronomique, par exemple, peut goûter plusieurs entrées ou plats principaux pour juger de la qualité et de la variété d'un menu. Les reporters des stations de télévision sont, en outre, souvent à la recherche de soi-disant « interviews de gens de la rue » pour déterminer comment la population perçoit un enjeu ou une question. Dans ces deux cas, on choisit l'échantillon au hasard, sans utiliser de méthode d'enquête particulière.

L'avantage évident de la méthode, c'est qu'elle est facile à utiliser, mais la présence de biais annule énormément ce dernier. Même si ses applications utiles sont limitées, la technique peut donner des résultats exacts lorsque la population est homogène.

Un scientifique pourrait, par exemple, utiliser cette méthode pour déterminer si un lac est pollué. En supposant que l'eau du lac est bien mélangée, tout échantillon donnerait de l'information identique. Un scientifique pourrait en toute sécurité puiser de l'eau n'importe où dans le lac, sans se tracasser au sujet de la représentativité de son échantillon.

Parmi les autres exemples d'échantillonnage de commodité, mentionnons :

- les femmes qui au cinéma s'assoient dans la première rangée;
- les 100 premiers clients à entrer dans un grand magasin;
- les trois premières personnes qui téléphonent à une station de radio dans le cadre d'un concours qu'elle a organisé.

Échantillonnage volontaire

Comme l'expression le laisse entendre, ce type d'échantillonnage intervient lorsque des gens offrent volontairement leurs services pour l'étude dont il est question. Il serait, par exemple, difficile et contraire à l'éthique dans le cadre d'expériences psychologiques ou d'essais de produits pharmaceutiques (de tests de médicaments) de recruter au hasard pour y participer des gens du grand public. En pareils cas, on prélève l'échantillon à partir d'un groupe de volontaires. Il arrive parfois qu'un chercheur offre de l'argent à des gens pour les inciter à participer à son étude. En échange, les volontaires acceptent la possibilité d'avoir à se prêter à des processus longs, exigeants ou quelques fois désagréables.

Le fait d'échantillonner des participants volontaires plutôt que la population en général peut introduire des biais marqués. Souvent, à l'occasion des sondages d'opinion, seuls les gens qui se soucient assez fortement d'une façon ou d'une autre de la question étudiée ont tendance à y répondre. La majorité silencieuse n'y répond généralement pas, ce qui entraîne un important biais sur le plan de la sélection.

Les stations de radio et de télévision ont souvent recours à des sondages par ligne ouverte pour interroger un auditoire ou un public sur ses vues. Le canal de télévision Much Music a recours à ce genre de sondages à l'intérieur de son émission CombatZone. Le canal demande durant cette dernière aux téléspectateurs de voter pour l'un des deux vidéoclips au sujet desquels il les invite à exprimer leur préférence en utilisant le téléphone, le courriel ou son site Web en ligne.

Bien souvent, on ne limite ni la fréquence ni le nombre des appels téléphoniques qu'un répondant peut effectuer en pareil cas. Une personne pourrait malheureusement, de ce fait, voter à plusieurs reprises. Il faut aussi noter que les gens qui participent à de tels sondages pourraient avoir des vues différentes de celles des gens qui ne le font pas.

Échantillonnage au jugé

On utilise la méthode d'échantillonnage au jugé lorsqu'on prélève un échantillon en se fondant sur certains jugements au sujet de l'ensemble de la population. L'hypothèse qui sous-tend son utilisation est que l'enquêteur sélectionnera des unités qui seront caractéristiques de la population. La question cruciale dans ce cas est l'objectivité : Dans quelle mesure peut-on se fier à son jugement pour en arriver à un échantillon typique? L'échantillonnage au jugé est exposé aux préjugés du chercheur et est peut-être encore davantage biaisé que l'échantillonnage de commodité ou à l'aveuglette. Étant donné que l'échantillonnage au jugé reflète toutes les idées préconçues que risque d'avoir le chercheur, il peut y avoir introduction de biais importants si ces idées sont inexactes.

Les statisticiens utilisent souvent cette méthode dans le cadre d'études préparatoires comme des tests préalables de questionnaires et des discussions en groupe. Ils préfèrent également avoir recours à cette méthode à l'intérieur du cadre de laboratoires où le choix des sujets des expériences (comme des animaux, des êtres humains et des végétaux) reflète les croyances ou les convictions antérieures de l'enquêteur au sujet de la population.

La réduction du coût et du temps qu'exige l'acquisition de l'échantillon est l'un des avantages de l'échantillonnage au jugé.

Échantillonnage par quotas

L'échantillonnage par quotas est l'une des formes les plus courantes d'échantillonnage non probabiliste. Il s'effectue jusqu'à ce qu'un nombre précis d'unités (de quotas) pour diverses sous-populations ait été sélectionné. Puisqu'il n'existe aucune règle qui régirait la façon dont il faudrait s'y prendre pour remplir ces quotas, l'*échantillonnage par quotas* est réellement un moyen de satisfaire aux objectifs en matière de taille d'échantillon pour certaines sous-populations.

Les quotas peuvent être fondés sur des proportions de la population. Si une population, par exemple, compte 100 hommes et 100 femmes et s'il faut en prélever un échantillon de 20 personnes pour qu'elles participent à un concours de dégustation de colas, il se peut que vous vouliez diviser l'échantillon en proportions égales entre les sexes, ce qui donnerait 10 hommes et 10 femmes. On peut penser que l'échantillonnage par quotas est préférable à d'autres formes d'échantillonnage non probabiliste (comme l'échantillonnage au jugé), parce qu'il impose l'inclusion dans l'échantillon de membres de différentes sous-populations.

L'échantillonnage par quotas est un peu similaire à l'échantillonnagestratifié parce que dans son cas également les unités semblables sont regroupées. Toutefois, il en diffère, cependant, sur le plan du mode de sélection. Dans le cas d'un échantillonnageprobabiliste, on sélectionne les unités au hasard, tandis que dans celui d'un échantillonnage par quotas, on laisse habituellement à l'intervieweur le soin de déterminer qui sera échantillonné. Cela peut donner lieu à des biais de sélection. Les responsables d'études de marché utilisent donc souvent l'échantillonnage par quotas (pour des enquêtes ou des sondages téléphoniques, en particulier), plutôt que l'échantillonnage stratifié, parce qu'il est relativement peu coûteux et facile à administrer et a la propriété souhaitable de respecter les proportions de la population. L'échantillonnage par quotas camoufle toutefois des biais pouvant être significatifs.

Comme dans le cas de toutes les autres méthodes d'échantillonnage non probabiliste, il faut supposer pour l'échantillonnage par quotas que les personnes sélectionnées sont semblables à celles qu'on ne sélectionne pas, afin de formuler des inférences au sujet de la population. Des hypothèses aussi audacieuses sont rarement valables.

Exemple n° 1 : Le conseil des élèves de l'école publique de la vallée de la rivière Rouge veut jauger l'opinion de ces derniers au sujet de la qualité de leurs activités parascolaires. Il décide d'interroger 100 des 1 000 élèves de l'école en utilisant comme sous-population les années d'études (c'est-à-dire les 7, 8, 9, 10, 11 et 12e années).

Le tableau ci-dessous fournit le nombre d'élèves.

Tableau 1. Nombre d'élèves inscrits à l'école publique de la vallée de la rivière Rouge, par année d'études

Année d'études	Nombre d'élèves	Pourcentage des élèves (%)	Quota d'élèves à l'intérieur de l'échantillon de 100
7	150	15	15
8	220	22	22
9	160	16	16
10	150	15	15
11	200	20	20
12	120	12	12
Total	**1 000**	**100**	**100**

Le conseil des élèves veut s'assurer que l'échantillon reflète le pourcentage d'élèves de chacune des années d'études. La formule est la suivante :

Pourcentage d'élèves en 10e année
= (nombre d'élèves de 10e année ÷ nombre total d'élèves) x 100 %
= (150 ÷ 1 000) x 100
= 15 %

Puisque 15 % des membres de la population de l'école sont en 10e année, l'échantillon devrait être constitué dans une proportion de 15 % d'élèves de 10e année. Utilisez, par conséquent, la formule suivante pour calculer le nombre d'élèves de 10e année qui devrait être inclus dans l'échantillon :

Échantillon d'élèves de 10e année
= (15 % de 100) x 100
= 0,15 x 100
= 15 élèves

La principale différence entre l'échantillonnage stratifié et l'échantillonnage par quotas tient au fait que le premier entraînerait la sélection des élèves à l'aide d'une méthode d'échantillonnage probabiliste comme l'échantillonnage aléatoire simple ou l'échantillonnage systématique. On n'utilise pas une telle technique dans le cas de l'échantillonnage par quotas. On pourrait sélectionner les 15 élèves en choisissant les 15 premiers élèves de 10e année qui entreraient à l'école une journée donnée ou en choisissant 15 élèves dans les deux premières rangées d'une classe en particulier. N'oubliez pas que les élèves qui arrivent en retard

ou qui s'assoient dans le fond de la classe peuvent avoir des opinions différentes de celles des élèves qui arrivent plus tôt à l'école ou qui s'assoient en avant lorsqu'ils entrent dans la classe.

Le fait que l'échantillonnage par quotas ne respecte pas l'exigence fondamentale du hasard est le principal argument militant contre son utilisation. Certaines unités peuvent n'avoir aucune chance d'être sélectionnées ou on risque de ne pas connaître leur chance de l'être. L'échantillon peut donc être biaisé.

Il est courant, mais il n'est pas nécessaire, que l'échantillonnage par quotas fasse appel à des procédures de sélection au hasard aux stades de départ, en grande partie de la même façon que le fait l'échantillonnage probabiliste. La première étape de l'échantillonnage à plusieurs degrés, par exemple, consisterait à sélectionner au hasard les régions géographiques. La différence se situe au niveau de la sélection des unités aux stades finals du processus.

Dans le cas de l'échantillonnageàplusieursdegrés, les unités reposent sur des listes à jour pour ce qui est des régions sélectionnées et on sélectionne un échantillon suivant un processus aléatoire. Dans le cas de l'échantillonnage par quotas, on indique à chaque intervieweur combien de répondants devraient être des hommes et combien d'entre eux, des femmes, de même que combien de gens devraient représenter les divers groupes d'âge. On calcule donc les quotas à partir des données dont on dispose pour la population; par conséquent, le sexe, les groupes d'âge ou d'autres variables démographiques sont représentés dans les bonnes proportions à l'intérieur des échantillons. Les intervieweurs peuvent, cependant, ne pas obtenir un échantillon représentatif de répondants dans le cas de chacun des quotas. Supposez, par exemple, qu'une organisation veut trouver de l'information sur les professions des hommes de 20 à 25 ans. Une intervieweuse se rend sur un campus d'université et sélectionne les 50 premiers hommes de 20 à 25 ans qu'elle rencontre par hasard et qui acceptent de participer à l'enquête de son organisation. Toutefois, cet échantillon ne signifie pas que ces 50 hommes sont représentatifs de tous les hommes de 20 à 25 ans.

L'échantillonnage par quotas est généralement moins coûteux que l'échantillonnage aléatoire. Il est également facile à administrer, compte tenu notamment du fait qu'on peut omettre de la procédure les tâches consistant à dresser la liste de la population entière, à sélectionner au hasard l'échantillon et à exercer un suivi auprès des non-répondants. L'échantillonnage par quotas, qui est une méthode d'échantillonnage efficace lorsqu'on a instamment besoin d'information, peut être effectué indépendamment des bases de sondage qui existent. Il peut être la seule méthode d'échantillonnage appropriée dans bien des cas où il n'existe pas de base de sondage convenable pour la population étudiée.

Estimation

Comme nous le savons maintenant, une enquête a pour but d'obtenir de l'information sur une population donnée. Lorsque l'échantillon a été sélectionné et que les données ont été recueillies (voir le chapitre Collectededonnées) et traitées (voir le chapitre Traitementdesdonnées), il reste à s'acquitter de la tâche consistant à apparier les données rassemblées à partir de l'échantillon à l'ensemble de la population.

L'estimation est le processus consistant à déterminer une valeur vraisemblable pour une variable de la population observée, en se fondant sur des données recueillies à partir de l'échantillon. Les chercheurs sont habituellement intéressés à examiner pour différentes variables des estimations d'un grand nombre de statistiques—des totaux, des moyennes et des proportions le plus souvent. On pourrait, par exemple, utiliser une enquête-échantillon pour produire les statistiques suivantes : des estimations de la proportion de fumeurs chez les 15 à 24 ans, les gains moyens des hommes et des femmes titulaires d'un grade universitaire et le nombre total de voitures que possède l'ensemble de la population observée.

Ce qui sous-tend le processus d'estimation, c'est le poids d'échantillonnage d'une unité, qui indique le nombre d'unités incluses dans la population (y compris le poids d'échantillonnage lui-même) et qui sont représentées par cette unité échantillonnée. Le poids d'échantillonnage est l'inverse de la probabilité de sélection de l'unité.

- **Exemple n° 1 :** Supposez que la Ville de Québec a décidé d'octroyer des cartes d'abonnement d'autobus à des usagers de ces derniers pour faire la promotion de ses services de transport en commun. Elle sélectionne à cette fin un échantillon aléatoire simple de 10 personnes parmi les 30 passagers d'un autobus. Puisque l'échantillonnage aléatoire simple donne à chaque membre de la population (tous les passagers de l'autobus dans ce cas) une chance égale d'être sélectionné, chacun des passagers avait une chance sur trois de l'être. Cela se traduit en un poids d'échantillonnage de trois pour chaque unité sélectionnée, ce qui signifie que chaque personne incluse dans l'échantillon représente trois personnes faisant partie de la population : elle-même, plus deux autres personnes.

  Afin d'estimer ce poids d'échantillonnage, on pourrait prendre les données d'enquête pour les 10 passagers sélectionnés et les copier trois fois afin de créer une population artificielle de 30 passagers. On pourrait ensuite estimer des totaux, des moyennes ou des proportions pour la population réelle à l'aide des statistiques correspondantes calculées au moyen de la population artificielle. Cependant, les statisticiens d'enquête attribuent plutôt un poids d'échantillonnage à chaque unité incluse dans l'échantillon et tiennent compte de ce poids lorsqu'ils établissent des estimations.

  Si une personne incluse dans un échantillon (d'un poids d'échantillonnage de 18) avait les yeux bleus et les cheveux bruns, ce serait alors comme si 18 personnes au total faisant partie de la population avaient les yeux bleus et les cheveux bruns.

  Exemple n° 2 : Vous êtes en train de mener une enquête pour déterminer le nombre total de gens qui habitent votre rue et le nombre moyen de voitures que possède chaque ménage qui y vit. Vous décidez de sélectionner un échantillonsystématique de 5 ménages à partir des 20 que compte votre rue et entendez utiliser cet échantillon pour estimer les totaux que vous cherchez à calculer. Le tableau qui suit résume l'information que vous avez rassemblée durant vos interviews auprès des ménages échantillonnés :

Tableau 1. Échantillon de ménages de la rue des Érables

Numéro de ménage	Nombre de personnes	Nombre de voitures	Probabilité de sélection	Poids d'échantillonnage
1	1	0	1/4	4
2	4	2	1/4	4
3	2	1	1/4	4
4	2	1	1/4	4
5	3	2	1/4	4

- La probabilité de sélection de 1 sur 4 découle du fait que l'échantillonnage systématique accorde une chance égale d'être sélectionné à chaque ménage de votre rue. Le poids d'échantillonnage de 4 est simplement l'inverse de cette probabilité. Lorsque vous établissez des estimations, vous devez examiner les caractéristiques de chaque ménage échantillonné. Dans ce cas, vous déterminez que 4 ménages sur la population des 20 que compte votre rue ont les mêmes caractéristiques.

  Pour estimer le nombre total de personnes habitant votre rue, vous devez multiplier le nombre de personnes membres d'un ménage par le nombre de ménages inclus dans ce poids d'échantillonnage, puis additionner tous les chiffres finals. Par exemple, il y a 4 ménages d'une personne (représentés par le numéro de ménage 1), 4 ménages de quatre personnes, 8 ménages de deux personnes (4 ménages représentés par le numéro de ménage 3 et 4 ménages représentés par le numéro de ménage 4) et 4 ménages de trois personnes. Votre estimation du nombre total de personnes serait alors la suivante :

  Nombre estimatif de personnes habitant votre rue

= (4 x 1) + (4 x 4) + (8 x 2) + (4 x 3)
= 48 personnes

Vous devez procéder de la même façon pour estimer le nombre moyen de voitures par ménage. Établissez une estimation du nombre total de voitures que possèdent les ménages domiciliés dans votre rue, puis divisez cette estimation par le nombre réel de ménages que compte votre rue. Il y a, par exemple, 4 ménages qui ne possèdent pas de voiture (représentés par le numéro de ménage 1), 8 ménages qui en possèdent deux (représentés par le numéro de ménage 2 et le numéro de ménage 5) et 8 ménages en possédant chacun un (représentés par le numéro de ménage 3 et le numéro de ménage 4).

Nombre estimatif de voitures
= (4 x 0) + (8 x 2) + (8 x 1)
= 24 voitures

Moyenne estimative
= 24 ÷ 20
= 1,2 voiture par ménage

Plans autopondérés

Toutes les unités échantillonnées n'ont pas toujours le même poids d'échantillonnage. Certains plans donnent aux unités une probabilité inégale d'être sélectionnées, ce qui fait que des unités incluses dans le même échantillon ont des poids d'échantillonnage différents. Les réponses d'un ménage ou d'une entreprise à un questionnaire pourraient représenter celles de 200 unités de la population, tandis que les réponses d'un autre ménage ou d'une autre entreprise au même questionnaire pourraient ne représenter que 50 unités faisant partie de la population.

Lorsque chaque unité incluse dans l'échantillon a le même poids d'échantillonnage, on dit que le plan d'échantillonnage est autopondéré. Ce genre de plan épargne du temps et est pratique sur le plan opérationnel, pour des échantillons de grande taille en particulier. Parce que chaque unité a le même poids, on peut ne pas tenir compte de tels poids lorsqu'on estime des moyennes et des proportions. La moyenne établie pour l'échantillon donne une estimation appropriée de la moyenne pour l'ensemble de la population.

Les plans d'échantillonnagealéatoire simple et d'échantillonnagesystématique sont des exemples de plans autopondérés. Ils auraient pu, pour cette raison, faciliter les calculs dans le cas de l'exemple n° 2. Pour estimer le nombre moyen de voitures par ménage inclus dans la population, par exemple, nous aurions pu utiliser la même moyenne que celle employée à l'intérieur de l'échantillon. Les 5 ménages échantillonnés possèdent au total 6 voitures, ce qui donne une moyenne de 1,2 voiture par ménage. C'est le même résultat que celui obtenu à l'aide de la procédure de calcul d'un poids d'échantillonnage.

Ajustement des poids

On ajuste parfois les poids d'échantillonnage avant une estimation, et ce, pour deux raisons fondamentalement :

- Pour tenir compte des non-réponses à un questionnaire : L'utilisation de poids d'échantillonnage pour établir une estimation donne de bons résultats lorsque vous avez pu interviewer toutes les unités sélectionnées. Dans l'exemple n° 2, si deux des cinq ménages échantillonnés avaient refusé de répondre à votre questionnaire ou n'étaient pas disponibles au moment de l'enquête, vous n'auriez des réponses que pour trois ménages, ce qui ne représenterait que 12 des 20 ménages que compte votre rue. Les deux unités qui n'auraient pas répondu à votre questionnaire représenteraient quatre ménages chacune, ce qui signifie que nous n'aurions aucune information sur le nombre de personnes ou de voitures pour huit ménages domiciliés dans votre rue. Afin d'apporter un ajustement pour en tenir compte, les statisticiens d'enquête accroissent habituellement les poids des unités ayant répondu au questionnaire d'enquête de manière à prendre en considération la perte de représentativité causée par les non-réponses à ce dernier. L'objectif consisterait à n'utiliser que les trois unités pour lesquelles nous disposerions d'information, mais qui représenteraient quand même les 20 ménages que compte votre rue.

- Pour tenir compte de données externes : Nous connaissons parfois le total réel pour une ou plusieurs variables mesurées à l'intérieur d'un échantillon. Dans l'exemple n° 3 de la section portant sur échantillonnage probabiliste on a divisé en proportions égales la population des 1 000 meilleurs films d'horreur, c'est-à-dire en 500 films classiques et en 500 films modernes. Même si vous connaissiez ce total avant l'échantillonnage, vous avez décidé de sélectionner un échantillon aléatoire simple de 100 films, ce qui vous a donné finalement 77 films classiques et 23 films modernes. Chacun de ces films avait un poids de 10 (parce que vous avez sélectionné 1 titre de film sur 10). Si vous utilisiez les réponses découlant de l'enquête et le poids d'échantillonnage, votre échantillon représenterait une population de 770 films classiques et de 230 films modernes. Cela pourrait entraîner des estimations inexactes. L'une des solutions à ce problème consisterait à réduire le poids de chaque film classique échantillonné et à accroître celui de chacun des films modernes échantillonnés également pour que votre échantillon donne une estimation de 500 films classiques et de 500 films modernes, ce qui réduirait la distorsion entraînée par un « mauvais » échantillon.

Évidemment, la stratification par date de parution sur les écrans avant l'échantillonnage aurait réglé le problème. Dans bien des cas, cependant, nous avons des totaux au niveau de la population, mais nous ignorons l'attribut de chaque unité incluse dans la base de sondage. Nous savons, par exemple, à partir du Recensement de la population, combien d'hommes et de femmes il y a dans une localité donnée, mais tout ce dont nous disposons pour un échantillonnage, c'est d'une liste de ménages. Il serait donc impossible de stratifier notre population selon le sexe. On utilise souvent des projections démographiques suivant l'âge et le sexe pour chaque province dans le cadre d'enquêtes sociales afin d'ajuster les poids d'échantillonnage.

On emploie les poids ajustés pour tenir compte des non-réponses à un questionnaire d'enquête et/ou de dénombrements externes à des fins d'estimation de la même façon qu'on a employé le poids d'échantillonnage dans l'exemple n° 1.

Autres méthodes d'estimation

L'utilisation des poids pour extrapoler les résultats de l'échantillonnage n'est pas l'unique méthode d'estimation qui existe, mais c'est la plus simple et la seule dont nous traiterons. Il est, néanmoins, important de savoir qu'il existe d'autres méthodes pouvant mener à des estimations plus précises (comme l'utilisation de données auxiliaires). Le processus d'estimation doit tenir compte du plan d'échantillonnage qui a été utilisé. Les estimations qui en résulteraient pourraient, autrement, être sérieusement biaisées.

Estimation de l'erreur d'échantillonnage

Comme nous l'avons déjà mentionné, toutes les estimations calculées à partir d'échantillons sont exposées à ce qu'on appelle l'erreur d'échantillonnage, qui découle du fait qu'on n'a observé qu'une partie, au lieu de la totalité, de la population. Un échantillon différent aurait pu fournir des résultats différents également. L'erreur d'échantillonnage est le degré de variation qui existe entre les estimations établies à partir des différents échantillons possibles. (Comme il y a en gros 14 millions de combinaisons différentes de 6 numéros de 1 à 49, imaginez alors combien il y a de façons de sélectionner un échantillon de 25 000

ménages canadiens!). On ne connaît pas, évidemment, cette erreur d'échantillonnage, puisque nous devrions connaître la réponse pour chaque unité de la population afin de la calculer. On peut, cependant, l'estimer en utilisant les données d'enquête. L'ampleur de l'erreur d'échantillonnage dépend de bien des éléments, y compris de la méthode d'échantillonnage, de la méthode d'estimation, de la taille de l'échantillon et de la variabilité de la caractéristique estimée. C'est pourquoi chaque estimation d'un échantillon entraîne une erreur d'échantillonnage qui lui est propre. On devrait donc lisser cette erreur pour chaque estimation d'un total, d'une moyenne, d'une proportion, etc., fournie par une enquête.

Exemples d'estimations à l'aide d'un plan d'échantillonnage aléatoire simple

L'échantillonnagealéatoiresimple est la plus simple de toutes les méthodes d'échantillonnage. On a déjà beaucoup étudié l'estimation effectuée à l'aide de la méthode d'échantillonnage aléatoire simple. Il existe des formules toute simples permettant d'estimer l'erreur d'échantillonnage pour bien des statistiques lorsqu'on a recours à un plan d'AES, étant donné surtout qu'il s'agit d'un plan d'échantillonnage autopondéré. Nous présentons ici l'estimateur le plus courant pour établir une moyenne (arithmétique) et un total d'une population, sous l'échantillonnage aléatoire simple.

Estimation de la moyenne d'une population

Dans le cas de l'échantillonnagealéatoiresimple, l'estimation de la moyenne d'une population est identique à la moyenne de l'échantillon :

$$\hat{\bar{X}} = \frac{\sum x}{n}$$

où
x = une valeur observée,

$\hat{\bar{X}}$
= estimation de la moyenne de la population,
$\sum$
x = somme de toutes les valeurs x observées dans l'échantillon,
n = nombre d'observations effectuées à l'intérieur de l'échantillon.

Nota : On devrait utiliser x et n (minuscules) si l'on renvoie à une enquête-échantillon et **X** et **N** (majuscules) lorsqu'on renvoie à une population.

Si les résultats de l'échantillonnage ont été résumés à l'intérieur d'une table des fréquences, l'estimation de la moyenne d'une population est alors la même que la moyenne de l'échantillon. Ainsi,

$$\hat{\bar{X}} = \frac{\sum x f}{\sum f}$$

où
x = une valeur observée,
f = la fréquence de la valeur (le nombre de fois que cette valeur a été observée dans l'échantillon),

$\hat{\bar{X}}$
= estimation de la moyenne de la population,
$\sum$
xf = somme de toutes les valeurs xf observées (le produit des valeurs observées multiplié par sa fréquence) dans l'échantillon,
$\sum$
f = somme des fréquences enregistrées à l'intérieur de l'échantillon.

Exemple n° 2 : Un producteur agricole sélectionne au hasard 10 œufs à partir d'une grosse (12 douzaines) d'œufs (144 œufs) qu'il trouve dans son poulailler. Il pèse soigneusement chacun des 10 œufs. Les poids suivants ont été enregistrés en grammes :

0,75, 0,70, 0,55, 0,50, 0,60, 0,65, 0,75, 0,65, 0,75 et 0,50

Quel est le poids moyen de ces œufs?

Nous pouvons déterminer à l'aide de la formule figurant ci-dessus le poids moyen des 10 œufs sélectionnés :

$$\hat{\bar{X}} = \frac{\sum x}{n}$$
$$= \frac{6,4}{10}$$
$$= 0,64 \text{ grammes}$$

Estimation du total d'une population

Dans le cas d'un échantillonnagealéatoiresimple, la formule d'estimation d'un total pour la population est :

$$\hat{X} = N \frac{\sum x}{n}$$

où
x = une valeur observée,

$\hat{X}$
= total estimatif de la population,
$\sum$
x = somme de toutes les valeurs x observées dans l'échantillon,
n = nombre d'observations effectuées dans l'échantillon,

N = nombre total d'observations effectuées à l'intérieur de la population.

C'est simplement l'estimation de la valeur moyenne multipliée par le nombre d'unités incluses dans la population. Dans l'exemple précédent, le poids moyen d'un œuf est 0,64 gramme; il est donc logique de penser que le poids total des 144 œufs serait 92,16 grammes (144 x 0,64 = 92,16 grammes).

Si les résultats de l'échantillonnage étaient résumés à l'intérieur d'une table des fréquences, la formule d'estimation pour la population totale serait :

$$\hat{x} = N \frac{\sum x f}{\sum f}$$

où
x = une valeur observée,

$\hat{x}$
= total estimatif de la population,
$\sum$
xf = somme de toutes les valeurs xf observées dans l'échantillon,
$\sum$
f = somme des fréquences enregistrées dans l'échantillon,
N = nombre total d'observations effectuées à l'intérieur de la population.

Exercices

1. Fait-on appel à l'échantillonnage aléatoire simple dans chacun des cas suivants? Expliquez brièvement comment on utilise ou non cette méthode d'échantillonnage dans chaque cas.
   a. une partie de bingo;
   b. une élection canadienne;
   c. un recensement.

2. Imaginez qu'une entreprise de fabrication de vêtements de la localité a 2 700 employés. Le directeur du personnel veut obtenir les suggestions des employés sur comment améliorer leur milieu de travail. Puisqu'il serait trop long de questionner tous les employés, le directeur choisit un échantillon systématique de 300 employés.
   a. Quel serait l'intervalle d'échantillonnage?
   b. Si le numéro 8 était votre premier numéro tiré au hasard, quels seraient les 5 premiers numéros de votre échantillon?

3. Supposez qu'une association sportive nationale mène une enquête auprès des gens vivant dans les capitales canadiennes pour savoir quel sport d'équipe ils préfèrent parmi les suivants : le hockey, la crosse, le curling et la ringuette.
   a. Copiez le tableau qui suit dans votre cahier. Inscrivez-y les données sur les provinces et les capitales qui manquent. (Vous pouvez obtenir les chiffres de population totale de 2006 à partir du site Web Profils des communautés de Statistique Canada. Dans l'espace « Nom de localité », tapez le nom de la capitale. Sélectionnez ensuite une province ou un territoire à partir de la liste déroulante. Cliquez sur le bouton « Recherche ». Lorsque les entrées disponibles apparaissent à l'écran, choisissez la 'ville' – p. ex. Québec (ville). N'utilisez pas les chiffres des agglomérations de recensement ni des régions métropolitaines de recensement [RMR].)

Tableau 1. Chiffres sur la population totale par capitale canadienne, 2006

Capitale	Territoire/province	Population totale Recensement de 2006
	Yukon	
Yellowknife		
	Nunavut	
Victoria	Colombie-Britannique	
	Alberta	
Regina		
Winnipeg	Manitoba	
Toronto		
Québec		
	Nouveau-Brunswick	
Halifax	Nouvelle-Écosse	
	Île-du-Prince-Édouard	
St. John's	Terre-Neuve et Labrador	

   b. Quelle est la population totale des capitales canadiennes combinées?

   c. Le tableau 2 fait état des résultats de l'enquête fictive. Examinez cette information et répondez aux questions suivantes :
      i. Dans quelle ville le hockey était-il le sport le plus populaire? le moins populaire?
      ii. Dans quelle ville la crosse était-elle le sport le plus populaire? le moins populaire?
      iii. Dans quelle ville le curling était-il le sport le plus populaire? le moins populaire?
      iv. Dans quelle ville la ringuette était-elle le sport le plus populaire? le moins populaire?

Tableau 2. Préférences pour les sports d'équipe, capitales canadiennes choisies

	Victoria	Fredericton	Regina	Yellowknife	Toronto	Québec
	%					
Hockey	10	3	58	23	71	29
Lacrosse	46	11	25	5	7	11
Curling	25	11	3	30	18	41

Ringuette	19	75	14	42	4	19
Total	**100**	**100**	**100**	**100**	**100**	**100**

d. Si vous deviez ventiler les données en d'autres catégories, que suggéreriez-vous?

4. 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 figurant ci-dessous indique le nombre d'élèves de l'école selon le sexe et l'année d'études, de la maternelle à la 12e année.

Tableau 3. Nombre d'élèves selon le sexe et l'année d'études, Académie Marguerite Bourgeois

	Maternelle	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**

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

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

c. 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 sexe masculin de la maternelle qu'il faudrait inclure dans l'échantillon.

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

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

e. 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.)

Activités en classe

1. Choisissez l'une des méthodes d'échantillonnage aléatoire décrites dans la présente section pour sonder les élèves de votre classe sur les sujets suivants :
   a. le nombre moyen d'enfants dans la famille;
   b. le mode de transport utilisé pour aller à l'école et en revenir;
   c. le nombre d'animaux familiers par ménage, leur type (c'est-à-dire un chat, un chien, etc.) et leur race.

2. Procurez-vous les registres des présences à deux cours donnés par votre titulaire sur lesquels sont indiqués le nom et le sexe de chaque élève. À l'aide de la technique d'échantillonnage stratifié, sondez 20 % des élèves ayant assisté à ces cours pour déterminer les sujets favoris de l'ensemble de ce groupe d'élèves. Les caractéristiques des strates seraient le sexe et le cours donné par votre titulaire.

Réponses

1. Voici une description du type de méthode d'échantillonnage utilisée dans chaque cas.
   a. Dans une partie de bingo, on utilise la méthode d'échantillonnage aléatoire simple. On place tous les numéros (la population totale) dans un baril et on en tire au hasard le nombre nécessaire (l'échantillon). Chaque unité a une chance égale d'être sélectionnée.

   b. Une élection canadienne est un exemple d'échantillonnage non aléatoire (d'échantillonnage volontaire), parce que chaque membre de la population (de 18 ans ou plus) peut y participer s'il le désire.

   c. Un recensement ne fait pas appel à l'échantillonnage aléatoire simple, parce qu'on doit y inclure tous les membres de la population cible.

2. Si l'on suppose que la taille de l'échantillon correspond à 300 personnes choisies d'une population de 2 700 employés,
   a. l'intervalle d'échantillonnage serait de 9 (2 700 ÷ 300 = 9).
   b. les 5 premiers numéros de l'échantillon seraient 8, 17, 26, 35 et 44.

3. Voici un tableau de la population totale de chacune des capitales provinciales et territoriales du Canada, selon le Recensement de la population de 2006.
   a.

Tableau 1. Population des capitales canadiennes, 2006

Capitale	Territoire/province	Population Recensement de 2006
Whitehorse	Yukon	*20 461*
Yellowknife	*Territoires du Nord-Ouest*	*18 700*
Iqaluit	Nunavut	*6 184*
Victoria	Colombie-Britannique	*78 057*
Edmonton	Alberta	*730 372*
Regina	*Saskatchewan*	*179 246*
Winnipeg	Manitoba	*633 451*
Toronto	Ontario	*2 503 281*
Québec	*Québec*	*491 142*
Fredericton	Nouveau-Brunswick	*50 535*
Halifax	Nouvelle-Écosse	*372 679*
Charlottetown	Île-du-Prince-Édouard	*32 174*
St. John's	Terre-Neuve et Labrador	*100 646*

Nota : Les entrées en italique sont les éléments du tableau qui manquaient dans la question 3. a).

   b. La population totale des capitales canadiennes serait de 5 216 928 personnes.
   c. L'information tirée du tableau 1 établit que :
      i. C'est à Toronto que le hockey était le plus populaire et à Fredericton qu'il était le moins populaire.
      ii. C'est à Victoria que la crosse était la plus populaire et à Yellowknife qu'elle était la moins populaire.
      iii. C'est à Québec que le curling était le plus populaire et à Regina qu'il était le moins populaire.
      iv. C'est à Fredericton que la ringuette était la plus populaire et à Toronto qu'elle était la moins populaire.

   d. On pourrait organiser les données selon le sexe et l'âge. On pourrait également demander aux répondants d'indiquer s'ils ont pratiqué le sport en question dans le passé ou s'ils le pratiquent actuellement ou s'ils en sont simplement des amateurs. On pourrait aussi poser d'autres questions du même genre.

4. Compte tenu de l'information fournie dans le tableau 3 de la question 4 :
   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 la maternelle 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 sexes par année d'études.

Tableau 3. Nombre d'élèves selon le sexe et l'année qu'il faut pour constituer un échantillon de 180 élèves de l'Académie Marguerite Bourgeois

	Maternelle	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**

Collecte de données

Collecteurs de données

Les particuliers et les organismes recueillent des données parce qu'ils ont besoin de se renseigner. Ils peuvent rechercher de l'information pour tenir des dossiers à des fins administratives, prendre des décisions au sujet d'importantes questions ou renseigner autrui. Quelle que soit la raison, on recueille des données pour informer.

Utilisateurs de l'information

Qui, dans la société, veut de l'information ou en a besoin? Voici certains des nombreux groupes et organismes qui utilisent les statistiques :

- **Administrations publiques :** Les administrations fédérale, 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, ainsi qu'à 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 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 :** La plupart des entreprises canadiennes ont besoin de renseignements sur l'activité économique locale 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.
- **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.
- **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.

On élabore souvent des statistiques par ce qu'on appelle communément une enquête. L'enquête statistique repose sur des concepts, des méthodes et des procédures bien définis et réunit les données recueillies en des structures de présentation utiles (comme des publications ou des articles de bulletin). Dans le cadre d'une enquête, on recueille différents types de données sur des sujets d'intérêt particulier. Les renseignements recueillis peuvent provenir de diverses unités d'une population (p. ex. un échantillon de téléspectateurs ou de gens qui font des courses le dimanche) ou de l'ensemble de ces unités (p. ex. les recensements de la population et de l'agriculture). Les données peuvent être recueillies soit directement auprès de la population échantillonnée, soit à partir de fichiers administratifs.

Les enquêtes menées auprès des populations (humaines ou non humaines) constituent une importante source de connaissances de base en sciences socioéconomiques. Nombreux sont les groupes d'intérêt — économistes, sociologues, etc. — qui obtiennent des subventions de l'État pour examiner des questions telles que la violence raciale dans les écoles, les comportements de l'électorat et le nombre d'enfants élevés dans des familles monoparentales.

Types de collecte de données

On peut recueillir les données à l'aide de trois grands types d'enquêtes, les recensements, les enquêtes-échantillon et les données administratives. Chacun présente à la fois des avantages et des inconvénients. Dans vos études, vous serez peut-être tenus de recueillir des données à certaines occasions. La méthode que vous choisirez à cette fin dépendra de divers facteurs.

Recensement

Par recensement, on entend la collecte de données sur tous les membres d'un groupe ou d'une population. Si vous aviez recueilli des données sur la taille de tous les élèves de votre classe, ce sera là un recensement de votre classe. Le recensement comporte aussi des « avantages » et des « inconvénients ». On peut préférer le recensement comme mode de collecte de données pour diverses raisons :

Avantages (+)

Variance d'échantillonnage nulle : Il n'y a pas de variabilité d'échantillonnage à attribuer aux statistiques parce qu'elles sont calculées à partir de données sur la population tout entière.

Niveau de détail : On peut établir des données se rapportant à de petites sous-populations.

Inconvénients (–)

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

Temps : Un recensement prend plus de temps à réaliser qu'une enquête-échantillon.

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

Contrôle : La tenue d'un recensement d'une grande population est une telle démarche qu'il est difficile de conserver un niveau de minutie et de contrôle qui soit le même à chaque étape.

Enquêtes-échantillon

Dans une enquête-échantillon, on recueille les données auprès d'une partie seulement de la population. Si les données obtenues portent sur la taille de 10 élèves dans une classe de 30, vous aurez utilisé une enquête-échantillon au lieu d'un recensement. Voici les raisons pour lesquelles on pourrait préférer ou non l'enquête-échantillon :

Avantages (+)

Coût : 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.

Temps : On obtient des résultats bien plus rapidement que dans un recensement. On communique avec moins d'unités, et il y a moins de données à traiter dans une enquête-échantillon que dans un recensement.

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

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

Inconvénients (–)

Variance d'échantillonnage non nulle : Il peut y avoir perte d'exactitude lorsque les données sont recueillies auprès d'une partie seulement d'un groupe plutôt qu'auprès de la population toute entière.

Niveau de détail : Il peut être impossible d'obtenir des renseignements sur une petite sous-population ou une petite région géographique.

Données administratives

Les organismes recueillent des données administratives dans le cadre de leurs activités quotidiennes, que ces données portent sur les naissances, les décès, les mariages, les divorces, ou les immatriculations de véhicules automobiles. Par exemple, avant de se voir délivrer un permis de mariage, un couple doit donner à l'officier de l'état civil des renseignements sur l'âge, le sexe, le lieu de naissance, les mariages antérieurs (s'il y en a) et le lieu de résidence. Ces données administratives peuvent être utilisées plus tard à titre de substitut pour une enquête-échantillon ou pour un recensement.

Avantages (+)

Variance d'échantillonnage nulle : Il n'y a pas de variabilité d'échantillonnage à attribuer aux statistiques parce qu'elles sont calculées à partir de données sur la population toute entière.

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

Simplicité : Avec des données administratives, il devient inutile de concevoir des activités de recensement ou d'enquête, ainsi que les travaux qui y sont liés.

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

Inconvénients (–)

Souplesse : À la différence des données d'enquête, les données administratives peuvent se limiter à des renseignements administratifs essentiels.

Population : Les données se limitent à la population figurant dans les dossiers administratifs.

Évolution au fil du temps : Les définitions, conçues à des fins précises, évoluent au fil du temps. Le statisticien doit s'attendre à ce que les définitions de ces dossiers évoluent.

Concepts et définitions : Ces définitions sont établies par ceux et celles qui conçoivent et gèrent le dossier selon leurs besoins. Par exemple, les définitions du revenu peuvent ne pas inclure tous les aspects prévus par les utilisateurs.

Qualité des données : L'accent mis sur la qualité des données peut varier d'un organisme à l'autre. Cela peut être évident lorsqu'une personne se fie aux données recueillies auprès d'un autre organisme.

Méthodes de collecte de données

Interviews sur place

Interview directe : Des intervieweurs qualifiés rendent visite à des gens pour recueillir des données. C'est un bon moyen pour assurer des taux de réponse élevés à un sondage ou à un recensement et, grâce aux intervieweurs, on recueille des données de meilleure qualité. Cette approche comporte toutefois des inconvénients : les répondants peuvent ne pas être disponibles pour une interview et le coût de déplacement des intervieweurs peut prendre des proportions considérables.

Interview sur place assistée par ordinateur : Il s'agit en quelque sorte d'un intervieweur sur place, sauf qu'au lieu d'avoir un questionnaire à remplir, l'intervieweur apporte avec lui un ordinateur portatif ou un ordinateur de poche pour entrer les renseignements directement dans la base de données. Cette méthode permet d'économiser du temps quant au traitement des données tout en évitant à l'intervieweur de transporter des centaines de questionnaires. Cependant, la préparation de ce type de méthode de collecte de données peut être coûteuse et requiert des intervieweurs qui possèdent des compétences en matière d'informatique et une habileté au clavier.

Interviews téléphoniques

Téléphone : Des intervieweurs qualifiés téléphonent aux répondants pour recueillir des données. Cette méthode d'enquête est plus rapide et moins coûteuse que l'interview directe. Il reste qu'on ne peut interroger que les gens disposant d'un téléphone, soit environ 98 % des Canadiens, et que le répondant a la possibilité de mettre abruptement fin à l'entretien.

Interview téléphonique assistée par ordinateur (ITAO) : C'est l'interview téléphonique classique, mais l'intervieweur entre les réponses des répondants directement dans l'ordinateur. Cela permet d'économiser du temps quant au traitement des données, mais la préparation de cette méthode peut être coûteuse et il faut des intervieweurs qui possèdent des compétences en matière d'informatique et une habileté au clavier. Statistique Canada mène des ITAO pour bon nombre de ces enquêtes, comme l'Enquête auprès des jeunes en transition, l'Enquête mensuelle sur les industries manufacturières, l'Enquête sociale générale et l'Enquête sur le lieu de travail et les employés.

Questionnaire à remplir soi-même

Enquête postale : Voici une méthode fréquente de collecte de données dans les enquêtes économiques de Statistique Canada. C'est un moyen de collecte relativement peu coûteux par lequel on peut distribuer un grand nombre de questionnaires en peu de temps. On a aussi la possibilité de communiquer avec des gens difficiles à joindre. Quant aux répondants, ils peuvent remplir le questionnaire au moment qui leur convient. Pour mener ce type d'enquête, il faut disposer d'une liste à jour de noms et d'adresses, ainsi qu'élaborer un questionnaire simple et limpide.

Les enquêtes postales ont comme principal inconvénient de recueillir des taux de réponse habituellement inférieurs à ceux d'autres méthodes de collecte de données, d'où d'éventuels problèmes de qualité des données statistiques. En outre, les gens 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.

Questionnaire livré en personne à domicile : Il s'agit d'enquêtes par autodénombrement où des questionnaires sont livrés à domicile et renvoyés par la poste une fois remplis par le répondant. Ce moyen permet habituellement d'obtenir de meilleurs taux de réponse que dans une enquête postale. Il convient tout particulièrement dans les cas où l'on a besoin de se renseigner auprès de plusieurs membres des ménages. Auparavant, Statistique Canada employait des questionnaires à livrer et à reprendre à domicile dans le cadre du Recensement de la population. Grâce aux questionnaires livrés à domicile et renvoyés par la poste par les répondants, on peut réduire les frais de collecte et sauvegarder davantage l'intimité des répondants inquiets du fait que quelqu'un se présente à leur domicile ou à leur entreprise pour recueillir les questionnaires.

Autres méthodes

Déclaration électronique des données (DED) : Depuis le début des années 1990, les formulaires électroniques de certaines enquêtes, principalement pour les enquêtes auprès des entreprises, étaient disponibles. Même si ce type de collecte de données est encore rare, il donne aux répondants l'option de choisir la formule qu'il préfère, soit de remplir le questionnaire habituel en format imprimé, soit utiliser la version électronique. L'évolution de la technologie est si rapide qu'il faut investir des sommes importantes pour maintenir des questionnaires protégés de qualité. Statistique Canada y met tous les efforts.

Internet : La popularité grandissante d'Internet a apporté des changements importants à la DED. Il est difficile de trouver une façon rapide et facile de recueillir des données dans Internet sans sacrifier les principes de Statistique Canada en ce qui a trait à la confidentialité, au respect de la vie privée et à la qualité des données. L'organisme a instauré des projets pilotes dans le cadre de diverses enquêtes importantes qui visent les répondants provenant des ménages, des universités, des entreprises et des ministères fédéraux. Depuis peu, les projets pilotes comprennent : le Recensement de la population de 2001, l'Enquête annuelle et l'Enquête sur le commerce de gros, l'Enquête unifiée auprès des entreprises et l'Enquête sur la rémunération auprès des entreprises.

On peut aussi procéder par observation directe, comme dans les enquêtes sur les prix ou consulter des dossiers administratifs existants. Le choix de la méthode dépend de divers facteurs : complexité et longueur du questionnaire, confidentialité des renseignements demandés, dispersion géographique de la population observée, coûts et délais.

Souvent, la stratégie de collecte la plus satisfaisante fera appel à une combinaison de méthodes. Ainsi, les enquêtes postales se sont révélées des plus efficaces lorsqu'elles sont assorties d'un suivi pour les personnes qui n'ont pas répondu au moyen d'une interview téléphonique.

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.

On doit considérer une foule de choses lorsqu'on entreprend d'élaborer 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 non-dirigées ou dirigées? Si elles sont dirigées, est-ce que les catégories de réponses sont mutuellement exclusives et 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 la nature de l'enquête (enquête à participation volontaire ou obligatoire) 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 comment les répondants peuvent avoir accès à ces données. 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 pn§sentent des avantages et des inconvenients. Une question ouverte permet au repondant d'interpreter ce qu'on lui dit et de repondre comme il le veut. Celui-ci donne sa reponse par ecrit au l'intervieweur inscrit mot a mot la reponse qu'il donne.

La question fermee oblige le repondant a choisir une reponse a partir des options de reponses. Pour le repondant, il est plus facile et plus rapide de repondre a des questions fermees. Dans le cas d'un chercheur, il est plus facile et mains coOteux de coder et d'analyser les reponses. En outre, les questions fermees produisent des reponses plus coherentes, ce qui n'est pas toujours le cas dans les questions ouvertes.

Essai du questionnaire

Il s'agit d'une etape fondamentale dans !'elaboration d'un questionnaire. L'essai du questionnaire aide a decouvrir des anomalies dans la formulation au l'ordre des questions; il permet de trouver les erreurs dans la presentation au les instructions d'un questionnaire; il permet aussi de cerner les problemes occasionnes par l'incapacite des repondants au leur refus de repondre aux questions; il permet de proposer des categories de reponses complementaires qui peuvent etre precodees dans le questionnaire et il fournit une indication provisoire de la duree de !'interview et des cas de refus. On peut effectuer un essai du questionnaire complet au d'une seule portion du questionnaire. A un moment donne, on devra cependant faire un essai complet du questionnaire.

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 mensuelles. 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'interviewer 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 préposé à la collecte des données sont les suivantes :

- posséder une bonne aptitude à communiquer;
- avoir une personnalité empreinte d'assurance et de professionnalisme;
- savoir utiliser l'automobile et le téléphone.

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 échantillonnée;
- donner au répondant le temps de lire ou de s'informer sur la confidentialité, l'enquête à participation volontaire ou obligatoire et l'échange de données avec d'autres organismes;
- Lire le message d'introduction du questionnaire au répondant (voir la section sur la conception des questionnaires).

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 libellées;
- ne pas suggérer de réponses au répondant;
- bien répondre aux questions du répondant;
- savoir garder le répondant informé;
- 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.

Exercices

Exercice sur la collecte de données

1. Après la Confédération, le premier recensement a eu lieu en 1871. Quelles sont les quatre provinces canadiennes fondatrices qui ont fait partie de ce recensement?
    a. Nouvelle-Écosse, Québec, Ontario et Manitoba
    b. Nouveau-Brunswick, Québec, Manitoba et Terre-Neuve
    c. Nouvelle-Écosse, Nouveau-Brunswick, Québec et Ontario
    d. Aucune des provinces énumérées ci-dessus

2. Lorsqu'on recueille des données, pourquoi est-il parfois préférable de mener une enquête-échantillon plutôt qu'un recensement?

3. Indiquez des raisons pour lesquelles on voudrait recueillir des données sur les sujets suivants :
    a. cambriolages
    b. causes de décès
    c. climat
    d. forêts
    e. immigration
    f. écoles

4. Quels éléments devriez-vous considérer au moment de choisir une méthode de collecte de données?

5. Compte tenu de certaines de vos réponses à la question 4, déterminez avec les élèves de votre classe quelle serait la méthode de collecte appropriée pour recueillir des données sur les sujets suivants :
    a. goûts musicaux des élèves de votre classe
    b. taille moyenne des élèves de votre classe
    c. temps que passent vos parents chaque semaine à faire des travaux ménagers
    d. attitude des élèves canadiens à l'égard de l'environnement

6. Y a-t-il des sujets sur lesquels on ne devrait pas recueillir de données, par exemple sur la santé ou les convictions politiques des gens? Discutez, avec les élèves de votre classe, de la limite à imposer pour déterminer si des sujets sont trop délicats pour faire l'objet de questions.

    Quels facteurs aideraient les chercheurs à décider? Par exemple, l'information que l'on veut recueillir est-elle d'importance nationale?

Exercice sur les questionnaires

Ce questionnaire a été conçu pour recueillir des renseignements sur les usagers de la bibliothèque municipale de Jolibois. Quelques problèmes se sont glissés dans ce questionnaire; veuillez les expliquer.

Nom :

Numéro de téléphone : Introduction

1. À quelle fréquence utilisez-vous les services offerts par la bibliothèque?

2. Combien de livres ou de périodiques avez-vous empruntés à la bibliothèque?

- 0
- 1 à 5
- 5 à 10
- 10 à 15
- 20 à 50
- 50 à 100

3. La dernière fois que vous vous êtes rendus à la bibliothèque, quel était le but de votre visite?

- recherche d'un livre
- recherche d'un périodique
- consultation auprès d'un bibliothécaire
- désir d'étudier en paix

4. Vos besoins ont-ils été comblés?

- Oui
- Non

5. Quel est votre degré de satisfaction à l'égard de la qualité du service offert par la bibliothèque et de l'attitude de son personnel?

1 2 3 4 5

6. Qu'est-ce qui vous déplaît à la bibliothèque?

7. Y a-t-il des améliorations qui pourraient être apportées à la bibliothèque en vue d'offrir un meilleur service?

- Oui
- Non

8. Êtes-vous en accord ou en désaccord avec les propositions récentes effectuées par le comité d'examen de la gestion de la bibliothèque, comme la recommandation de doubler les amendes pour les livres dont le délai d'emprunt est échu?

- En accord
- En désaccord

(Passez à la **question 11**.)

9. Êtes-vous au courant de ces propositions?

- Oui
- Non

10. Pourquoi êtes-vous en désaccord avec ces propositions?

11. Êtes-vous contre le fait que les heures d'ouverture de la bibliothèque ne soient pas plus longues?

- Contre
- Pas contre

12. Niveau de scolarité :

Merci de votre aide!

Réponses

Réponses à l'exercice sur la collecte des données

1. c) Nouvelle-Écosse, Nouveau-Brunswick, Québec et Ontario.

2. Il est parfois préférable de mener une enquête-échantillon plutôt qu'un recensement pour les raisons suivantes : cela coûte moins cher, se fait plus rapidement, répond à des besoins spécialisés et allège le fardeau du répondant. De plus, il suffit peut-être de recueillir des renseignements auprès d'un segment particulier de la population.

3.

4. 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

Discutez des autres questions avec les élèves de votre classe.

Réponses à l'exercice sur les questionnaires

Voici en bref quelques problèmes que pourrait poser le questionnaire de la bibliothèque municipale de Julibois. En avez-vous relevé d'autres? Discutez-en avec les élèves de la classe.

Introduction

Il manque un mot de présentation où l'on décrit le but de l'enquête et l'usage qui sera fait des résultats. Cette introduction permettrait au lecteur qui n'est pas un usager de la bibliothèque de « s'auto-éliminer » et de ne pas répondre au questionnaire.

Il est préférable de ne pas demander le nom et le numéro de téléphone si l'on veut obtenir des réponses honnêtes et ouvertes. Les gens aiment savoir que les renseignements qu'ils fournissent sont confidentiels et le resteront.

Question 1 :

Il s'agit d'une question « ouverte » qui serait mieux formulée comme question « fermée » qui permettrait de comparer les réponses. En outre, on devrait préciser la période de référence (c'est-à-dire combien de fois a-t-on fréquenté la bibliothèque au cours de la dernière semaine, ou du dernier mois, ou des 12 derniers mois).

Question 2 :

La question porte sur les emprunts de livres ou de périodiques. Qu'en est-il si l'on a emprunté les deux à la fois? Quelle réponse cocheriez-vous si vous aviez emprunté cinq livres et cinq périodiques? Choisiriez-vous « 1 à 5 » ou « 5 à 10 »? L'autre problème, c'est qu'on ne définit pas la période de référence. A-t-on fait ces emprunts lors du dernier mois, de la dernière année ou des 10 dernières années?

Question 3 :

C'est une question fermée qui fournit seulement quatre choix de réponse. On peut avoir d'autres raisons de fréquenter la bibliothèque. On pourrait ajouter une catégorie « Autre » accompagnée d'un champ « Veuillez préciser » ou une liste plus longue de raisons (incluant « activités sociales » et « séances de signature »). En outre, on pourrait ajouter au début de la question la directive « Indiquez toutes les raisons qui s'appliquent ». De cette façon, le répondant peut cocher plus d'un choix.

Question 4 :

La question convient telle qu'elle est posée mais ne va pas assez loin. Les responsables de la bibliothèque voudront probablement savoir pour quelles raisons les besoins n'ont pas été comblés. Cette question devrait comporter une deuxième partie qui demande au répondant d'expliquer comment ses besoins n'ont pas été comblés.

Question 5 :

Il s'agit d'une « question double » à laquelle le répondant a deux réponses à fournir dans une. Un autre problème a trait aux choix de réponses : on ne donne aucune explication sur la signification des chiffres. Le répondant doit-il évaluer son degré de satisfaction au moyen des chiffres? Si oui, que représente le chiffre 1? Est-ce que le chiffre 5 signifie mauvais ou excellent?

Question 6 :

C'est une question ouverte pour laquelle il sera difficile de coder et de totaliser les différentes réponses possibles. Pour simplifier la tâche aux répondants et aux analystes des données, on pourrait fournir une liste de réponses possibles à cocher et ajouter une dernière catégorie « Autre » en demandant de préciser.

Question 7 :

La réponse à cette question est « Oui » ou « Non ». Toutefois, le répondant peut ne pas être informé au sujet des améliorations à apporter à la bibliothèque ou ne pas avoir d'opinion.

Questions 8, 9 et 10 :

La question 9 (« Êtes-vous au courant de ces propositions? ») devrait figurer avant la question 8 (« Êtes-vous en accord ou en désaccord avec les propositions récentes … »), car la question 8 n'est pertinente que si l'on a répondu « Oui » à la question 9.

À la question 8, on demande au répondant d'approuver ou de désapprouver toutes les propositions. Qu'en est-il s'il en accepte quelques-unes et rejette les autres? Les choix de réponse présentés ne donnent pas une véritable mesure des réactions aux propositions. La formulation de la question 8 devrait être modifiée afin d'éviter les biais de réponse.

La partie « Passer à » de la question 8 constitue également un problème. Elle renvoie le répondant à la question 11 et, par conséquent, tous les répondants qui ont indiqué « En désaccord » à la question 8 ne pourraient pas répondre aux questions 9 et 10. Ainsi, on devrait lire « (Passer à la question 10.) ».

Question 11 :

Voici une formulation en double négation. Toutes sortes d'interprétations sont possibles. On pourrait reformuler la question plus simplement : « Êtes-vous en faveur de la prolongation des heures d'ouverture de la bibliothèque? » en présentant des choix de réponse à cocher tels que « Oui », « Non » et « Ça m'est indifférent ».

Question 12 :

C'est une question délicate à laquelle certains répondants peuvent ne pas être à l'aise de répondre ou répondre honnêtement. Il serait peut-être préférable de ne pas poser cette question. Toutefois, si le niveau de scolarité est un renseignement pertinent et nécessaire, une question fermée se prêterait davantage à une analyse quantitative.

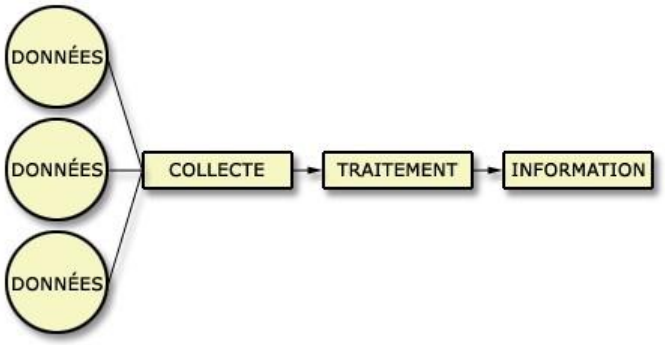
Traitement des données

Les données sont des faits ou des nombres qui permettent de tirer des conclusions. Une fois enregistrées, classées et organisées, les données deviennent de l'information lorsqu'elles sont liées à un cadre qui leur donne une signification ou qu'elles sont interprétées en fonction de ce cadre. La conversion des données en information compte plusieurs étapes, que nous désignerons ici comme les étapes de traitement des données. Le présent chapitre porte sur le traitement des données et la manière dont les ordinateurs effectuent ces étapes rapidement et facilement.

Introduction

L'organigramme simplifié ci-dessous montre la transformation des données brutes en information. Le traitement des données a lieu après la collecte de toutes les données pertinentes auprès de diverses sources et leur saisie dans un ordinateur qui en fera le traitement pour produire l'information (résultats).

Figure 1. Organigramme du traitement des données



Le traitement des données regroupe les étapes suivantes :

- Codage
- Saisie
- Vérification
- Imputation
- Contrôle de la qualité
- Production des résultats

Codage

D'abord les données brutes doivent être codées avant leur saisie dans un ordinateur. Pour ce faire, les réponses obtenues lors des enquêtes doivent être libellées, habituellement au moyen de codes numériques simples. Cette étape peut faire partie des tâches de l'intervieweur sur le terrain ou même d'un employé de bureau. Le codage des données est important, parce qu'il facilite la saisie et le traitement des données.

Les enquêtes comptent deux types de questions — *les questions fermées* et *les questions ouvertes*. Les réponses à ces questions influent sur le type de codage réalisé. Une question fermée signifie qu'il y a un nombre fixe de réponses prédéterminées. Ces réponses ont déjà été codées.

La question suivante posée lors de l'Enquête de 1998 sur l'emploi du temps (Sport) est un exemple d'une question fermée :

Jusqu'à quel point le sport est-il important pour vous procurer les avantages suivants?

<1/> Très important

<2/> Quelque peu important

<3/> Pas important

Une question ouverte signifie que toutes les réponses sont permises, compliquant ainsi le codage. Pour coder une question ouverte, le processeur doit échantillonner un nombre de réponses, puis concevoir une structure de codes qui inclut toutes les réponses possibles.

La structure de codes ci-après est un exemple d'une question ouverte :

Quels sports pratiquez-vous?

Veuillez préciser <28 caractères/>_____

Le codage est déjà inscrit dans les champs de réponse des questionnaires utilisés aux fins du Recensement et de la plupart des autres enquêtes. Lorsque le temps est venu de faire le dépouillement des questionnaires, les codes sont entrés directement dans la base de données afin de pouvoir en faire la saisie. En voici un exemple :

Quelle langue cette personne parle-t-elle le plus souvent à la maison?

<18/> Anglais

<19/> Français

<20/> Autre — Veuillez préciser _____

Statistique Canada a élaboré des codes uniformes, désignés classifications ou normes, pour aider à répartir les personnes, les endroits et les choses en groupes spécialisés. Se reporter à la section sur la classification pour obtenir plus de renseignements à ce sujet.

Systèmes de codage automatisés

Statistique Canada teste constamment de nouveaux programmes destinés à automatiser les tâches répétitives et routinières. Le codage se prête donc parfaitement à ce type d'automatisation.

Parmi les avantages que confère un système de codage automatisé dans le processus de traitement, notons entre autres :

- la rapidité;
- l'uniformité;
- l'économie.

Nous utilisons déjà certains systèmes automatisés. Par exemple, les fichiers de données de l'Enquête sur la population active sont recueillis par les bureaux régionaux de Statistique Canada et passent dans un système de codage automatisé qui attribue les codes des branches d'activité et des professions en fonction du Système de classification type des industries et du Système de classification type des professions. Les dossiers rejetés (ceux qui ne concordent pas avec la réponse écrite) sont les seules données qui font l'objet d'un codage manuel.

Lors du Recensement de la population de 1991, on a eu recours à un système de codage automatisé pour coder la langue, la religion et l'origine ethnique. Ces codes ne font partie d'aucun système de classification uniformisée, mais pour le recensement, ils tiennent lieu de norme qui facilite l'analyse et la tabulation.

La prochaine étape du traitement des données consiste à saisir les données codées dans une base de données informatisée. Cette étape s'appelle la saisie de données.

Saisie des données

Il s'agit du processus par lequel on transfère les données d'une copie papier (questionnaires et réponses d'enquête) vers un fichier électronique. Les réponses sont ensuite transmises à un ordinateur.

Avant cette procédure, il importe de *nettoyer* (préparer) les questionnaires en vue de la saisie. À cette étape, une personne examine le questionnaire pour veiller à la présence de toutes les données pertinentes minimales et à leur lisibilité. Habituellement, ce nettoyage se fait lors de vérifications automatiques exhaustives.

Il existe plusieurs méthodes de saisie :

- **Les tableaux de dépouillement** servent à enregistrer les données, comme le nombre de fois que survient un événement particulier, et à élaborer des tableaux de répartition de la fréquence. Le tableau de dépouillement est utilisé seulement en présence d'une petite quantité de données. Cette forme de saisie ne serait jamais utilisée pour un recensement. Toutefois, lorsqu'il est nécessaire de procéder à un décompte manuel (p. ex., bulletins de vote), un tableau de dépouillement se révélerait très utile.

- **La saisie par lots** constitue l'une des plus anciennes méthodes de saisie. Elle recourt à un clavier pour entrer les données. Ce processus convient parfaitement à l'entrée d'une grande quantité de données lorsque la rapidité est une nécessité. Aucune procédure de vérification n'a lieu, mais nous devons placer une bonne part de la confiance dans le programme de vérification. En outre, nous devons mettre en oeuvre des vérifications de la validité et des étendues pour veiller à la qualité de la saisie. Cela ne signifie pas que nous modifions les données, mais si un champ numérique contient plutôt des lettres, l'erreur sera détectée. Cette approche peut être avantageuse dans le cadre de grandes enquêtes comportant de nombreuses questions et vérifications.

- **La saisie interactive** est souvent désignée *saisie intelligente*. En général, les données saisies font l'objet d'une vérification avant leur imputation. Toutefois, cette méthode combine la saisie et la vérification en une seule fonction. Bien que la saisie interactive soit plus lente, elle est très efficace lorsqu'il y a une étroite interdépendance entre les questions. Ce processus nécessite une connaissance des procédures de vérification, car il importe de corriger les erreurs sur le champ. La saisie interactive réduit aussi le nombre de documents traités, car les vérifications ont lieu directement sur un ordinateur.

  Les vérifications réalisées durant la saisie interactive peuvent prendre la forme de vérifications préliminaires des données ou de vérifications élémentaires (comme la vérification des totaux, des sous-totaux et des valeurs), ou de vérifications complètes appliquées à l'ensemble du questionnaire. (Un de ces systèmes est désigné Interview téléphonique assistée par ordinateur et fait l'objet du chapitre Collecte des données).

- **Le lecteur de reconnaissance** ou le lecteur de code à barres peut reconnaître des lettres et des chiffres. Ces lecteurs balaient chaque ligne et les convertissent dans le programme.

- Le lecteur de code à barres est très répandu, surtout dans les magasins. Il peut se présenter sous la forme d'un pistolet ou d'une baguette. Le lecteur en forme de pistolet est simple à programmer et peut vérifier la validité des caractères lus. Le Recensement de l'agriculture utilise cette méthode de lecture dans son système de commande des documents. Une imprimante imprime les codes à barres pour libeller les questionnaires. Ensuite, le lecteur lit l'étiquette, convertit les lignes en nombres et les enregistre dans un fichier de commande de document.

  Le lecteur de type baguette ou crayon lit les numéros d'identification. Cette méthode de lecture nécessite une main lente et sûre pour bien aligner la baguette avec les caractères imprimés.

- **L'enregistrement magnétique** permet la lecture et l'écriture. Cette méthode est utilisée lorsqu'il importe de garantir la sécurité des données. L'application la plus répandue de ce type de saisie est la lecture du NIP des cartes bancaires.

Le clavier d'un ordinateur est le dispositif d'entrée le plus connu actuellement. Auparavant, les employés saisissaient les données au moyen de cartes perforées ou d'un ruban papier.

Voici quelques-uns des dispositifs modernes de saisie :

- le lecteur optique de marques
- le lecteur de code à barres
- le numériseur utilisé pour l'éditique
- le crayon optique
- la boule roulante
- la souris

Saviez-vous que...

- En 2001, cette étape visait en outre la création d'un système d'extraction d'images donnant accès aux images de tous les questionnaires et de tous les registres des visites, de façon à éliminer la manutention de milliers de boîtes et de documents papier dans le cadre des processus subséquents nécessitant la consultation de l'original des questionnaires du recensement, comme lors des recensements antérieurs.

- Les utilisateurs peuvent obtenir sans frais davantage de données du recensement de 2001 dans Internet en accédant au site Web de Statistique Canada.

- En 2001, le recensement recueille pour la première fois des renseignements sur les couples formés de partenaires de même sexe, ainsi que de l'information sur la langue parlée au travail.

- Comme on n'imprime plus le guide aux répondants qui accompagnait auparavant le questionnaire abrégé, quelque 215 460 kg de papier sont épargnés.

- Environ 13,2 millions de questionnaires seront saisis par clavier, ce qui équivaut à 5 milliards de frappes.
- Parmi les autres outils utilisés pour la collecte et le traitement des données, il y l'Internet qui permet aux Canadiens de remplir leur questionnaire en ligne, le processus de contrôle automatisé, l'application d'interview téléphonique assistée par ordinateur et le système connexe de gestion des dossiers;

Après la saisie des données dans la base de données informatisées, l'étape suivante consiste à vérifier l'exactitude de toutes les réponses. Cette méthode est connue sous le nom de « *vérification des données* ».

Vérification des données

Il importe de vérifier les données avant de les présenter à titre d'information. Nous assurons ainsi l'exactitude, l'exhaustivité et la cohérence de l'information. Peu importe le type de données utilisées, toutes les enquêtes font l'objet de certaines vérifications. Cette vérification peut se faire manuellement, avec l'aide d'un programme informatique, ou en combinant ces deux méthodes, selon le support (électronique ou papier) sur lequel se trouvent les données.

Il existe deux niveaux de vérification — la *microvérification* et la *macrovérification*.

La microvérification consiste à corriger les données au niveau de l'enregistrement. Ce processus vise à détecter les erreurs par la vérification de chaque enregistrement de données. Il a pour but de déterminer la cohérence des données et de corriger chaque enregistrement.

La macrovérification vise aussi à détecter les erreurs, mais par l'analyse des données agrégées (totaux). Les données sont comparées à celles des autres enquêtes, des fichiers administratifs ou de versions antérieures des mêmes données. Ce processus détermine la compatibilité des données.

Nous pourrions poser la question « Pourquoi avons-nous des erreurs dans nos fichiers? » Des erreurs peuvent se glisser dans les données pour toutes sortes de raisons. Voici quelques exemples :

- Un répondant peut avoir mal interprété une question.
- Un répondant ou un intervieweur peut avoir coché la mauvaise réponse. Un
- intervieweur peut avoir mal codé ou mal interprété une réponse écrite. Un
- intervieweur peut avoir omis une question ou oublié d'écrire la réponse. Un
- répondant peut avoir donné des réponses inexactes.

On doit toujours tenir compte des objectifs de la vérification de manière à :

- assurer l'exactitude des données; assurer la
- cohérence des données; déterminer
- l'exhaustivité des données; assurer la
- cohérence des données agrégées; obtenir les
- meilleures données possibles.

Application des règles de vérification

Alors, comment cette vérification se fait-elle? La première étape consiste à appliquer aux données des « règles » (ou des facteurs dont il faut tenir compte). Ces règles résultent de la connaissance experte d'un spécialiste du domaine, de la structure du questionnaire, de l'historique des données et d'autres enquêtes ou données connexes.

La connaissance experte peut provenir de diverses sources. Le spécialiste peut être un analyste qui possède une vaste expérience du type de données vérifiées. Un expert peut aussi être un commanditaire de l'enquête, familier avec les relations entre les données.

La disposition et la structure du questionnaire influent aussi sur les règles de vérification. Par exemple, parfois les répondants doivent sauter certaines questions si elles ne s'appliquent pas à eux ou à leur situation. Il importe de respecter cette spécification et de l'intégrer dans les règles de vérification.

Enfin, d'autres enquêtes portant sur le même type de variables ou de caractéristiques servent à établir certaines règles de vérification.

Types de vérifications

Il existe plusieurs types de vérifications, dont :

- la **vérification de la validité**, qui porte sur un champ ou une cellule à la fois. Elle fait en sorte que nous tenions compte des descripteurs d'enregistrement, des caractères et des valeurs invalides; que les champs obligatoires aient été remplis (p. ex., aucun champ de quantité ne demeure vierge si un nombre est requis); que nous ayons bien utilisé les unités de mesure précisées et que le temps de déclaration respecte les limites précisées.
- la **vérification de l'étendue** ressemble à la vérification de la validité, car elle porte aussi sur un champ à la fois. Ce type de vérification a pour but d'assurer que les valeurs, les rapports et les calculs respectent les limites prédéterminées.
- la **vérification des dédoublements** examine un enregistrement complet à la fois. Elle recherche les dédoublements de sorte que l'enregistrement d'une personne ou d'une chose n'apparaît qu'une seule fois. Cette vérification recherche aussi si le répondant apparaît plus d'une fois dans l'univers de l'enquête, surtout après la modification du nom. Enfin, elle vérifie aussi si les données ont été entrées une seule fois dans le système.
- la **vérification de la cohérence** compare les diverses réponses d'un enregistrement pour en assurer la cohérence l'une par rapport à l'autre. Par exemple, si une personne déclare faire partie du groupe d'âge de 0 à 14 ans et est à la retraite, il existe alors un problème de cohérence entre les deux réponses. La vérification inter-champ constitue une autre forme de vérification de la cohérence. Elle vise à vérifier que, si un chiffre est déclaré dans une section, un chiffre équivalent sera déclaré dans une autre section.
- la **vérification historique** sert à comparer les réponses dans les enquêtes régulières (c.-à-d., détecter toute modification importante par rapport à la déclaration précédente). Elle compare aussi les rapports et les calculs. Ainsi, tout écart qui ne respecte pas les limites établies est signalé et fait l'objet

d'une analyse.

- la **vérification statistique** compare un ensemble complet de données. Ce type de vérification a lieu seulement après l'exécution de toutes les autres vérifications et la correction subséquente des données. Les données sont compilées et toute valeur extrême, toute donnée suspecte et toute valeur aberrante est rejetée.

- les **vérifications diverses** portent sur les dispositions spéciales de déclaration, les vérifications dynamiques propres à l'enquête, les vérifications de la classification appropriée, les modifications apportées aux adresses physiques, aux lieux et aux contacts ainsi que la vérification de la lisibilité, qui permet de s'assurer que les chiffres et les symboles sont reconnaissables.

La vérification des données dépend de la complexité du questionnaire. Cette complexité vise la longueur et le nombre de questions posées. Elle porte aussi sur les détails des questions et l'étendue du domaine que couvre le questionnaire. Dans certains cas, la question peut utiliser une terminologie très technique. Dans ce type d'enquêtes, nous aurons recours à des méthodes spéciales de déclaration et à des vérifications propres à cette industrie.

Données erronées

La vérification des données devrait détecter et minimiser les erreurs, telles que :

- les questions non posées
- les réponses non enregistrées
- les réponses non pertinentes

Une réponse inexacte peut résulter d'une négligence ou d'un effort délibéré de donner des réponses déroutantes. Elle aura aussi lieu lorsque la réponse nécessite un calcul mathématique. Par exemple, la conversion des jours en heures ou d'un revenu annuel en revenu hebdomadaire accroît le risque d'erreurs.

Exemple 1 - Réponses inexactes

Cet exemple de vérification illustre la manière dont peut se glisser une réponse inexacte. Lire attentivement les questions et réponses suivantes, tirées du formulaire de l'Enquête sur la population active de Statistique Canada. Où se trouve l'erreur dans la réponse du répondant?

Question 151 - Outre le temps supplémentaire, combien d'heures rémunérées la personne n° 1 a-t-elle travaillées chaque semaine?
Réponse - 40

Question 153 - La semaine dernière, pendant combien d'heures la personne n° 1 a été absente de son emploi en raison de vacances, de maladie ou de toute autre raison?
Réponse - 0

Question 155 - La semaine dernière, combien d'heures supplémentaires rémunérées la personne n° 1 a-t-elle travaillées à son emploi?
Réponse - 4

Question 156 - La semaine dernière, combien d'heures supplémentaires non rémunérées la personne n° 1 a-t-elle travaillées à son emploi?
Réponse - 0

Question 157 - La semaine dernière, combien d'heures la personne n° 1 a-t-elle réellement travaillées à son emploi principal?
Réponse - 40

La question 151 indique que la personne 1 travaille habituellement 40 heures par semaine. La question 153 montre que le répondant n'a pas eu de temps libre la semaine précédente, et la question 155 indique en fait que la personne a travaillé quelques heures supplémentaires. Toutefois, la question 157 précise que la personne a travaillé au total 40 heures durant la semaine! La réponse réelle devrait plutôt être 44 heures.

Les réponses à chaque question semblent acceptables. C'est seulement lors de leur comparaison que nous pouvons relever des inexactitudes dans une ou plusieurs des réponses données.

Le recoupement, type de vérification de la cohérence, est seulement un type de vérification. Il permet la comparaison des réponses données à diverses questions. Nous pouvons l'effectuer manuellement ou au moyen d'un logiciel de vérification.

Cette vérification indique que, dans l'exemple précédent, il est nécessaire de prendre d'autres mesures pour obtenir une réponse exacte, l'intervieweur devra communiquer de nouveau avec le ménage et vérifier le nombre d'heures travaillées par la personne 1.

Lors d'une interview sur place ou téléphonique assistée par ordinateur, l'intervieweur aurait aussitôt obtenu un avertissement électronique à la saisie de 40 en réponse à la question 157. L'intervieweur aurait alors pu contre-vérifier la réponse auprès du répondant. Cette méthode est beaucoup plus rapide et évite le fardeau de recommuniquer avec le répondant.

Vérification : outil de gestion

La vérification est aussi un outil précieux pour évaluer la qualité des données, car elle peut indiquer s'il est nécessaire d'apporter de nombreuses vérifications. Elle permet aussi d'éviter la tenue d'une autre enquête semblable, car elle permet de relever la cause probable des problèmes.

Imputation

En elle-même, la vérification apporte peu à l'amélioration générale des résultats réels de l'enquête si aucune mesure corrective n'est prise lorsque les articles ne respectent pas les règles établies durant la vérification. Lorsqu'il manque des données à un fichier, l'imputation a lieu en général après le traitement des données en fonction des règles de vérification.

Les non-réponses et les données invalides influent définitivement sur la qualité des résultats de l'enquête. L'imputation résout les problèmes des réponses manquantes, invalides ou incomplètes relevées durant la vérification, ainsi que les erreurs de vérification qui peuvent survenir. À ce stade, toutes les données font l'objet d'une recherche d'erreurs parce que les répondants ne sont pas les seuls à commettre des erreurs, qui risquent aussi de se glisser durant le codage et la vérification.

Les procédures d'imputation visent à combler les lacunes. Ainsi, les modifications sont apportées au nombre minimal de champs jusqu'à ce que les enregistrements aient fait l'objet de toutes les vérifications. À la détection de ces erreurs, les données des entrées invalides, manquantes ou incomplètes sont imputées ou remplacées par des valeurs appropriées. Nous attribuons aussi des réponses aux questions non répondues. Cette procédure relève de personnes qui ont un accès complet aux microdonnées et qui détiennent de bons renseignements auxiliaires.

Les procédures d'imputation sont établies lors de la planification et de l'élaboration de l'enquête. Certains problèmes sont éliminés plus tôt par une communication avec le répondant ou par l'examen manuel du questionnaire, mais en général il est impossible de résoudre tous les problèmes en raison du fardeau de réponse, du coût et des contraintes de temps. Ainsi, l'imputation sert à combler les lacunes de la vérification.

Bien que l'imputation puisse améliorer la qualité des données finales, il importe de choisir avec soin la méthodologie d'imputation pertinente. Certaines méthodes d'imputation ne préservent pas la relation entre les variables. En fait, certaines introduisent même une distorsion dans la répartition sous-jacente.

L'imputation des données exige que nous tenions compte de plusieurs facteurs. En général, l'imputation déductive est la première méthode utilisée. Cette méthode est utilisée lorsqu'on peut déduire une valeur avec certitude. Elle peut avoir lieu durant la collecte, la saisie, la vérification ou les étapes ultérieures du traitement des données. L'imputation déductive sert lorsqu'il n'y a qu'une seule réponse possible à la question (p. ex., toutes les valeurs sont données mais il manque le total ou le sous-total).

Voici d'autres méthodes d'imputation :

- La **méthode du hot deck** recourt à d'autres enregistrements (donneurs) de la même enquête pour répondre à la question (ou à l'ensemble de questions) qui doit faire l'objet d'une imputation. Le donneur peut être choisi au hasard parmi un groupe de donneurs affichant le même jeu de caractéristiques prédéterminées. Par exemple, dans le cas où un questionnaire serait retourné parce qu'il manquait le revenu dans un enregistrement, nous pourrions regrouper toutes les unités déclarantes de la même province, de la même profession et du même nombre d'années d'expérience, puis en choisir un au hasard. On dresserait ensuite une liste des donneurs qui répondent à ces critères, puis on en choisirait un au hasard. Le revenu déclaré par cette personne tiendrait lieu de revenu pour la réponse manquante ou invalide.

- La **méthode de la substitution** repose sur la disponibilité de données comparables. Les données imputées peuvent être extraites de l'enregistrement du répondant obtenu lors d'un cycle antérieur de l'enquête ou d'un autre fichier source (p. ex. fichiers administratifs ou autres fichiers d'enquête du même répondant). Cette méthode est souvent difficile à réaliser parce que, dans de nombreux cas, il n'existe aucune autre information au sujet de ce répondant, à part l'information de l'enquête en cours.

- La **méthode de l'estimateur** recourt à une information tirée d'une autre question ou d'une autre réponse d'un enregistrement qui fera l'objet d'une imputation (du cycle actuel ou d'un cycle antérieur) et, au moyen d'opérations mathématiques, produit une valeur plausible destinée au champ vierge ou erroné.

  La méthode la plus simple est désignée *imputation moyenne*. Cette méthode permet l'insertion dans le champ vierge d'une valeur moyenne tirée des unités déclarantes affichant le même jeu de caractéristiques prédéterminées. Par exemple, s'il manque le revenu à un enregistrement, nous pourrions imputer le revenu moyen de la même province pour la même profession et la même expérience. Il existe aussi d'autres méthodes de l'estimateur plus poussées.
- La **méthode du cold deck** ressemble à celle du hot deck, sauf que le donneur ne provient pas de la même enquête. Il peut venir d'enquêtes antérieures ou alors d'un recensement. Ces valeurs peuvent découler de données historiques, de l'expertise de la spécialité, etc. Un questionnaire « parfait » est créé pour répondre aux besoins complets ou partiels de l'imputation.

- On peut aussi trouver le donneur par une méthode appelée **estimateur du plus proche voisin**. Dans un tel cas, nous devons élaborer, à partir de caractéristiques prédéterminées, un critère quelconque pour déterminer laquelle des unités déclarantes ressemble le plus à l'unité pour laquelle il manque une valeur. L'unité la plus semblable tient alors lieu de donneur.

La méthode d'imputation peut varier d'une enquête à l'autre et, dans des circonstances uniques ou particulières, à l'intérieur d'une même enquête. Ces méthodes peuvent être appliquées manuellement ou au moyen d'un système automatisé. La valeur imputée est déterminée à la suite d'une communication avec le répondant ou par le jugement d'un spécialiste du sujet. Pour faciliter l'imputation, Statistique Canada a écrit des programmes spécialisés d'imputation de données à la suite de commentaires sur la méthodologie formulés par des statisticiens d'expérience qui ont analysé l'enquête et suggéré des approches pour imputer avec la plus grande justesse des données significatives.

Les méthodes d'imputation peuvent être effectuées automatiquement, manuellement ou en combinaison. Bien réalisée, l'imputation restreint les écarts causés par l'absence d'enregistrements complets et précis, offre une piste de vérification destinée à l'évaluation et veille à uniformiser à l'interne les enregistrements imputés. Une bonne procédure d'imputation est automatisée, objective et efficiente.

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 à l'objectif visé 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 oeuvre, 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;
- 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;
- la réaction aux différences.

Par exemple, notons le contrôle de la qualité des activités de codage, d'interviewing 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.

Le tableau suivant fait état des principales distinctions entre l'assurance 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 pré-établie pour permettre la comparaison
- s'applique principalement à l'étape du traitement
- est une procédure sous-jacente à l'assurance de la qualité

Gestion de la qualité dans les organismes statistiques

La qualité des données doit être établie et assurée dans le contexte de la pertinence. La fonction prévue des données et les caractéristiques fondamentales de la qualité influent sur la pertinence des données et des renseignements statistiques. Nous devons aussi tenir compte du type de données que l'utilisateur juge utile.

Dans les organismes statistiques, il n'existe aucune définition uniforme des *statistiques officielles*. Il existe bien des facteurs de qualité reconnus, mais évolutifs, sous-jacents au concept de la pertinence. Nous devons tenir compte de ces facteurs de qualité et les équilibrer lors de la conception et de la mise en oeuvre du programme statistique de l'organisme.

Donc, comment pouvons-nous définir la qualité? Voyons la liste des facteurs de qualité énoncée ci-dessous :

- La **pertinence** de l'information statistique reflète la mesure dans laquelle cette information répond aux besoins réels des clients. C'est en examinant cet élément qu'on détermine si l'information disponible permet de mieux comprendre les enjeux qui sont importants pour les utilisateurs. L'évaluation de la pertinence est subjective et est fonction des différents besoins des utilisateurs. Le défi du Bureau consiste à mettre en balance les besoins conflictuels des utilisateurs actuels et potentiels afin d'élaborer un programme qui répond aux besoins les plus importants dans un contexte de ressources limitées.

- L'**exactitude** de l'information statistique est la mesure dans laquelle l'information décrit correctement le phénomène qu'elle devrait évaluer. Généralement, elle est caractérisée par l'erreur dans les estimations statistiques et est décomposée en composantes de biais (erreur systématique) et de variance (erreur aléatoire). L'exactitude peut également être décrite en fonction des sources d'erreur majeures qui peuvent mener à l'inexactitude (p. ex., couverture, échantillonnage, non-réponse, réponse).

- L'**actualité** de l'information statistique renvoie à l'intervalle entre le point de référence (ou la fin de la période de référence) auquel se rapporte l'information et la date à laquelle l'information est diffusée. Habituellement, l'actualité se trouve en relation d'équilibre avec l'exactitude. L'actualité de l'information influera sur sa pertinence.

- L'**accessibilité** de l'information statistique renvoie à la facilité avec laquelle on peut obtenir l'information auprès du Bureau. Cela comprend la facilité avec laquelle on peut certifier l'existence de l'information, ainsi que l'à-propos de la forme ou du médium par le biais duquel on peut accéder à l'information. En outre, le coût de l'information peut représenter un facteur d'accessibilité pour certains utilisateurs.

- L'**intelligibilité** de l'information statistique reflète la disponibilité de l'information et des métadonnées supplémentaires nécessaires à l'inteprétation et à l'utilisation appropriées des renseignements. Normalement, cette information comprend les variables, les classifications et les concepts sous-jacents utilisés, la méthode de collecte des données et le traitement, ainsi que les indications ou les mesures de l'exactitude de l'information statistique.

- La **cohérence** de l'information statistique reflète la mesure dans laquelle on peut réussir à regrouper cette information avec d'autres renseignements statistiques dans un cadre analytique général et au fil du temps. L'utilisation de classifications, des populations cibles et de concepts normalisés favorise la cohérence, tout comme l'utilisation d'une méthode commune d'une enquête à l'autre. La cohérence ne sous-tend pas nécessairement l'uniformisation numérique complète.

Ces facteurs de qualité tendent à se chevaucher, souvent de manière confuse. Tout comme il n'existe aucune mesure simple de l'exactitude, il n'existe aucun modèle statistique efficace qui regroupe tous ces facteurs de qualité en un seul indicateur. De même, sauf dans les cas simples et unidimensionnels, il n'existe aucun modèle statistique général qui permet de déterminer si un jeu particulier de facteurs offre une meilleure qualité qu'un autre.

L'atteinte d'un niveau de qualité acceptable résulte de la prise en considération, de la gestion et de l'équilibre au fil du temps des divers facteurs ou éléments qui constitue une meilleur qualité. Il en est ainsi de l'attention portée aux objectifs du programme, aux principales utilisations des données, aux coûts de même qu'aux conditions et aux circonstances qui influent sur la qualité des données et les attentes de l'utilisateur. Étant donné que les facteurs de qualité entretiennent des liens complexes, une mesure prise pour corriger ou modifier un des aspects de la qualité tend à influer sur les autres facteurs. En conséquence, l'équilibre entre ces facteurs peut être rompu de manière imprévisible ou non quantifiable au départ. La décision et les mesures qui visent à atteindre cet équilibre reposent sur la connaissance, l'expérience, les analyses, les commentaires, la consultation et, inévitablement, le jugement.

Production des résultats

Après la vérification, nous pouvons traiter davantage les données pour produire les résultats escomptés. Le logiciel utilisé pour traiter les données dépend du format de présentation nécessaire.

Les logiciels de traitement de texte, d'éditique, de production de diagrammes (comprenant des diagrammes et des dessins), de programmation, de base de données et de tabulation sont régulièrement utilisés. Voici quelques exemples de logiciels de présentation des données :

- Les **tableurs** sont des programmes qui additionnent automatiquement des colonnes et des lignes de nombres, calculent des moyennes et réalisent des analyses statistiques. On peut s'en servir pour produire des tableaux sommaires des résultats. Les tableurs peuvent aussi produire des diagrammes qui illustrent la signification d'un groupe de données. Ceux-ci se présentent sous une diversité de formes : diagrammes à barres, diagrammes linéaires simples, diagrammes à secteurs, pour ne nommer que quelques exemples des données visuelles que nous pouvons produire.
- Les **bases de données sont des classeurs électroniques**. Ils enregistrent systématiquement les données pour en faciliter l'accès et en produire des sommaires, des agrégats et des 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 finaux publiables.

Les résultats peuvent servir de nombreuses façons. Ils peuvent être enregistrés en vue d'une récupération et d'une utilisation ultérieures. Ils peuvent être imprimés sur papier sous forme de tableaux ou de diagrammes ou même copiés sur des transparents destinés à un rétroprojecteur. Les résultats peuvent aussi être enregistrés sur support électronique afin de pouvoir les utiliser sur des ordinateurs portatifs ou personnels. Enfin, les résultats peuvent être transmis par Internet à d'autres personnes sous forme de fichiers électroniques.

Les résultats sont habituellement conçus dans le but de répondre au besoin de communiquer une information particulière à un auditoire précis. La seule limite aux diverses formes de résultats produits tient aux divers types de dispositifs de sortie actuellement offerts. Pour aider à choisir le meilleur type de sortie pour l'information produite, il suffit de se poser les questions suivantes : À qui sont destinés les résultats produits? Sous quel format les résultats seront-ils le mieux compris?

Exercice

1. D'après ce que vous savez déjà du processus statistique, placez les étapes suivantes dans le bon ordre :
   - traitement
   - collecte
   - information
   - données

2. Énumérez toutes les étapes nécessaires au traitement des données et décrivez-les brièvement.

3. Recherchez les divers types de dispositifs de saisie disponibles à votre école. L'utilisation de ces dispositifs nécessite-t-elle des compétences spéciales? Discutez-en en classe.

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

5. Les prochaines questions montrent quelques exemples de réponses d'enquête incorrectes. Trouvez les erreurs de chaque exemple et expliquez la raison de leur inexactitude.

   a.

Question 5a

Question	Réponse
(i) **Combien y a-t-il de pièces dans ce logement?** **Inclure** cuisine, chambres à coucher, pièces finies dans le grenier ou le sous-sol, etc. **Exclure** salles de bains, halls, vestibules et pièces utilisées uniquement pour le travail.	11
(ii) **Combien de ces pièces sont des chambres à coucher?**	Oui

   b.

Question 5b

Question	Réponse
Quel est votre état matrimonial actuel? *Cochez une **seule** réponse.*	• ( ) Marié(e) • ( ) Séparé(e) • (x) Divorcé(e) • ( ) Veuf(ve) • (x) Jamais marié(e) (célibataire)

   c.

Question 5c

Question	Réponse
De quelle manière vous êtes-vous rendu au travail le jour du Recensement 2001? *Si vous avez utilisé plusieurs moyens de transport, cochez toutes les réponses pertinentes.*	• ( ) Automobile, camion ou camionnette — à titre de conducteur • ( ) Automobile, camion ou camionnette — à titre de passager • ( ) Transport public (p.ex., autobus, tramway, métro, train léger, train de banlieue, traversier) • ( ) Marche • ( ) Bicyclette • (x) Motocyclette • ( ) Taxi • ( ) Autre moyen • (x) Travaille à la maison • (x) N'a pas travaillé

6. Repondez aux questions suivantes :
    - Votre ecole a-t-elle recourt aux procedures de codage pour produire les donnees?
    - Les procedures de codage sont-elles effectuees par une personne ou une machine?
    - Si les procedures de codage sont manuelles, serait-il preferable de les automatiser? Pourquoi?

Réponses

1. Le bon ordre de ces étapes est le suivant :
    - données
    - collecte
    - traitement
    - information

2. Les étapes sont le codage, la saisie, la vérification, l'imputation, la qualité et la production des résultats.

3.

4. 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.

5.
    a. L'erreur dans ce questionnaire est imputable à la réponse du répondant. Le répondant aurait dû entrer le nombre de chambres à coucher à la partie (ii), et non la réponse « oui ».
    b. Cet exemple comporte deux erreurs. D'abord, la question demande au répondant de ne cocher qu'une seule réponse. Le répondant a plutôt coché deux réponses. Ensuite, le répondant a coché à la fois « divorcé(e) » et « jamais marié(e) ». Une personne ne peut pas être divorcée sans s'être auparavant mariée.
    c. Dans cet exemple, le répondant prétend qu'il est allé au travail en motocyclette. Mais le répondant a aussi coché la réponse « n'a pas travaillé ». Si la personne n'a pas travaillé, elle ne peut pas avoir voyagé en motocyclette pour se rendre au travail.

    Toutefois, le répondant aurait pu légitimement cocher les réponses « motocyclette » et « travaille à la maison », parce qu'il aurait pu conduire jusque chez lui (lieu de travail possible) à partir d'un autre endroit.

Organisation des données

Une fois la collecte et le traitement terminés, les données doivent être organisées de façon à produire de l'information utile. Il est bon de se familiariser avec certaines définitions quand on organise des données. Le présent chapitre fournit un aperçu de ces définitions et certaines techniques simples pour organiser et présenter des données.

Variables

Comme on utilise souvent le mot **variable** dans le cadre de l'étude de la statistique, il est important de comprendre sa signification. Une variable est une caractéristique qui peut supposer plus d'un ensemble de valeurs auquel il est possible d'attribuer une mesure numérique.

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. On peut classer les variables en diverses catégories, dont certaines sont brièvement définies dans la présente section.

Variables catégoriques

Une variable catégorique (aussi appelée variable qualitative) est une variable où chaque réponse peut être classée dans une catégorie particulière. Ces catégories doivent être mutuellement exclusives et exhaustives. Une catégorie mutuellement exclusive signifie que toutes les réponses d'enquête possibles doivent faire partie d'une seule catégorie, alors qu'une catégorie exhaustive signifie qu'une catégorie doit tenir compte de toutes les réponses possibles. Une variable catégorique peut être nominale ou ordinale.

Variables nominales

Une **variable nominale** décrit un nom ou une catégorie. Contrairement à une variable ordinale, les noms ou les catégories possibles ne suivent pas un ordre naturel. Le sexe et le genre de logement en sont des exemples. Dans le tableau 1, la variable « Mode de transport pour se rendre au travail » est une variable nominale, parce qu'elle décrit une catégorie de mode de transport.

Tableau 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

Source : Recensement de 2001 (Série « analyses ») N° au catalogue : 96F0030XIF2001010

Variables ordinales

Une **variable ordinale** est une variable dite catégorique dans laquelle les catégories possibles peuvent être classées dans un ordre spécifique ou dans un ordre naturel quelconque. Dans le tableau 2, la variable « comportement » est ordinale parce que la catégorie « Excellent » est meilleure que la catégorie « Très bon », etc. On n'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 ».

Tableau 2. Classement des élèves selon le comportement

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

Variables numériques

Une **variable numérique**, connue aussi sous le nom de variable quantitative, est une variable qui peut supposer un nombre infini de nombres réels, comme l'âge ou le nombre de membres d'un ménage. Toutefois, on ne considère pas que toutes les variables décrites par des nombres sont des variables numériques. Par exemple, lorsqu'on vous demande d'indiquer votre niveau de satisfaction par une valeur allant de 1 à 5, vous utilisez des nombres, alors que la variable « satisfaction » est en fait une variable ordinale. Les variables numériques peuvent être continues ou discrètes.

Variables continues

On dit qu'une variable est continue si elle peut supposer un nombre infini de valeurs réelles. La distance, l'âge et la température sont des exemples d'une **variable continue**. Les méthodes utilisées ou la précision des instruments employés pour mesurer une variable continue en restreignent la mesure. La taille d'un élève, par exemple, est une variable continue, parce qu'elle peut être de 1,6321748755... mètre (m).

Quand on mesure la taille d'une personne, cependant, on la calcule habituellement en l'arrondissant au centimètre près. Par conséquent, on enregistrerait la taille de l'élève susmentionné à 1,63 m.

Nota : Pour en faciliter la manipulation, on groupe habituellement les variables continues à l'aide d'<< intervalles de classe >>, dont il sera question plus loin dans la presente section. Parce qu'ils font partie du processus d'organisation des donnees, les groupements de variables deviennent de l'information utile.

Variables discretes

Contrairement à une variable continue, une **variable discrete** ne peut revetir qu'un nombre defini de valeurs reelles. La note accordee par un juge à un gymnaste à une competition est un exemple de variable discrete : la plage varie de 0 à 10 et la note ne comporte jamais plus qu'une decimale (P.-Jl., une note de 8,5). On peut aussi grouper des variables discretes. Encore une fois, on groupe des variables discretes pour en faciliter la manipulation.

Nota : La mesure d'une variable continue est toujours une approximation discrete.

Tableaux de distribution de fréquences

La *fréquence* **(f)** d'une observation particulière est le nombre de fois que celle-ci se dégage des données. La *distribution* d'une variable est le profil des valeurs de l'observation. Les distributions de fréquences sont représentées sous forme de tableaux, d'histogrammes ou de polygonesdefréquences.

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, que nous expliquerons un peu plus loin.

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 1. Tableau de distribution de fréquences du nombredevéhiculesimmatriculésdanslesménages

Nombre de véhicules (x)	Comptage	Fréquence (f)
0	𝍸	4
1	𝍸 I	6
2	𝍸	5
3	III	3
4	II	2

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, 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 (c.-à-d., 1) à la fréquence de l'intervalle 45 à 54 (c.-à-d., 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 :

Tableau2.Âgesdesparticipantsàuntournoid'échecs

Valeur inférieure	Valeur supérieure	Fréquence (f)	Fréquence cumulée	Pourcentage	Pourcentage cumulé
35	44	1	1	10,0	10,0
45	54	2	3	20,0	30,0
55	64	2	5	20,0	50,0
65	74	2	7	20,0	70,0
75	84	2	9	20,0	90,0
85	94	1	10	10,0	100,0

Voir la section « Fréquencecumulée et pourcentagecumulé » pour plus de renseignements sur la façon de construire des tableaux de fréquences cumulées.

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, 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 :

Tableau3.DuréedeviedespilesAA,enminutes

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

Fréquence relative et fréquence en pourcentage

Un analyste qui étudierait ces données 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

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

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 :

Tableau4.DuréedeviedespilesAA,enminutes

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	0,10	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	0,10	10
430 à 439	1	0,03	3
Total	**30**	**1,00**	**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.

N'oubliez pas que ces énoncés statistiques supposent qu'il y a eu prélèvement d'un échantillon représentatif. En réalité, un analyste se reporterait aussi à une estimation de la variabilité (voir la section intitulée Mesuresdedispersion) pour compléter l'analyse. Toutefois, pour nos besoins, il suffit de savoir que les tableaux de distribution de fréquences peuvent fournir de l'information importante sur la population pour laquelle un échantillon a été prélevé.

Diagrammes à tiges et à feuilles

Un diagramme à tiges et à feuilles est une technique utilisée pour classer des variablesdiscrètes ou continues. Il sert à organiser des données au fur et à mesure de leur collecte.

Un diagramme à tiges et à feuilles ressemble un peu à un diagramme à bandes. Chaque nombre à l'intérieur des données est ventilé en une tige et une feuille, d'où le nom diagramme à tiges et à feuilles. La tige du nombre inclut tous les chiffres sauf le dernier. La feuille du nombre sera toujours un chiffre unique.

Éléments d'un bon diagramme à tiges et à feuilles

- Un bon diagramme à tiges et à feuilles indique comme tige les premiers chiffres du nombre (des milliers, des centaines ou des dizaines) et comme feuille le dernier chiffre (unique).
- Un diagramme à tiges et à feuilles indique habituellement des nombres entiers. Un nombre avec un point décimal sera arrondi au nombre entier le plus près. Par exemple, des résultats d'examens, des mesures de taille et de poids, des données de vitesse, etc.
- Un bon diagramme à tiges et à feuilles ressemble à un diagramme à bandes quand on le tourne de côté.
- Il indique comment les données sont dispersées, c'est-à-dire le nombre le plus élevé, le nombre le plus faible, le nombre le plus courant ou le plus fréquent et les valeurs aberrantes (un nombre qui se situe à l'extérieur du groupe principal de nombres).

Tuyaux sur la façon de dessiner un diagramme à tiges et à feuilles

Une fois que vous aurez décidé qu'un diagramme à tiges et à feuilles est le meilleur moyen d'illustrer vos données, dessinez-le comme suit :

- Inscrivez les milliers, les centaines ou les dizaines (tous les chiffres, sauf le dernier) dans la partie gauche de la page. Ces chiffres constitueront les tiges.
- Tracez une ligne à la droite des tiges.
- De l'autre côté de la ligne, inscrivez les chiffres (le dernier chiffre d'un nombre) qui constitueront les feuilles.

Par exemple, si la valeur observée est 25, la tige sera alors 2 et la feuille, 5. Par ailleurs, si la valeur observée est 369, la tige sera 36 et la feuille, 9. Dans les cas où les observations comportent une décimale ou plus, telle que 23,7, la tige sera 23, alors que la feuille sera 7. Si la plage des valeurs est trop grande, il suffit d'arrondir le chiffre 23,7 à 24, de façon à limiter le nombre de tiges.

Les coches ne sont pas nécessaires lorsqu'on dessine un diagramme à tiges et à feuilles, puisqu'on utilise des données réelles.

Pas si simple à comprendre? Faites quelques exercices.

Exemple 1 – Création d'un diagramme à tiges et à feuilles

Chaque matin, une enseignante interrogeait dans le cadre d'un jeu-questionnaire les élèves de sa classe en leur posant 20 questions au sujet de la géographie. Les élèves consignaient les notes de la classe et leurs propres notes également. À mesure que l'année avançait, chaque élève essayait d'améliorer ses notes au jeu-questionnaire. Samuel enregistrait ses notes au jeu-questionnaire chaque matin sur un diagramme à tiges et à feuilles. Voici à quoi elles ressemblaient tracées sous forme de points :

Tableau 1. Notes obtenues par Samuel l'an dernier au jeu-questionnaire sur des faits bien connus

Tige	Feuille
0	3 6 5
1	0 1 4 3 5 6 5 6 8 9 7 9
2	0 0 0 0

Analysez le diagramme à tiges et à feuilles de Samuel. Quelle a été sa note la plus courante au jeu-questionnaire portant sur la géographie? Quelle a été sa note la plus élevée? Sa note la plus faible? Tournez de côté le diagramme à tiges et à feuilles pour qu'il ressemble à un diagramme à bandes. Est-ce que la plupart des notes de Samuel se situent dans les 10, les 20 ou les moins que 10? Comme on ne connait pas l'ordre des notes dans le tracé, c'est difficile de déterminer si Samuel s'est amélioré ou non.

Essayez de créer votre propre diagramme à tiges et à feuilles. Utilisez toutes les notes d'examen que vous avez obtenues l'année dernière ou les points accumulés par votre équipe sportive au cours de la saison.

Le principal avantage d'un diagramme à tiges et à feuilles

Le principal avantage d'un diagramme à tiges et à feuilles, c'est que les données sont groupées et que toutes les données originales y sont indiquées également. Dans l'exemple3 sur la durée de vie des piles de la section des Tableaux de distribution de fréquences, un tableau montre que deux observations font partie de l'intervalle variant de 360 à 369 minutes. Le tableau ne vous dit pas toutefois en quoi consistent réellement ces observations. Un diagramme à tiges et à feuilles permettrait d'illustrer cette information. Sans diagramme à tiges et à feuilles, on ne peut trouver les deux valeurs (363 et 369) qu'en parcourant toutes les données originales, une tâche fastidieuse lorsqu'on a beaucoup de données!

Quand on examine un ensemble de données, on peut considérer que chaque observation se compose de deux parties : une tige et une feuille. Pour créer un diagramme à tiges et à feuilles, il faut d'abord séparer chaque valeur observée en deux parties :

- La tige est le premier chiffre ou la première série de chiffres;
- La feuille est le dernier chiffre d'une valeur;
- Chaque tige peut se composer d'un certain nombre de chiffres;
- Chaque feuille ne peut comporter qu'un seul chiffre.

Exemple 2 – Création d'un diagramme à tiges et à feuilles

Un enseignant a demandé à 10 élèves d'indiquer combien de livres ils avaient lus au cours des 12 derniers mois. Voici leurs réponses :

12, 23, 19, 6, 10, 7, 15, 25, 21, 12

Tracez un diagramme à tiges et à feuilles pour ces données.

Tuyau : On peut écrire le chiffre 6 comme 06 pour désigner que la tige est 0 et la feuille, 6.

Voici ce à quoi le diagramme à tiges et à feuilles devrait ressembler :

Tableau 2. Livres lus en une année par 10 élèves

Tige	Feuille
0	6 7
1	2 9 0 5 2
2	3 5 1

Dans le tableau 2 :

- la tige 0 représente l'intervalle de classe 0 à 9;
- la tige 1 représente l'intervalle de classe 10 à 19;
- la tige 2 représente l'intervalle de classe 20 à 29.

Habituellement, un diagramme à tiges et à feuilles est ordonné, ce qui veut dire simplement que les feuilles sont disposées dans l'ordre ascendant de gauche à droite. De plus, les signes de ponctuation (les virgules ou les points) qui séparent les feuilles (chiffres) ne sont pas nécessaires, puisque chaque feuille est toujours un chiffre unique.

Nous avons créé, à l'aide des données tirées du tableau 2, le diagramme à tiges et à feuilles qui figure ci-dessous :

Tableau 3. Livres lus en une année par 10 élèves

Tige	Feuille
0	6 7
1	0 2 2 5 9
2	1 3 5

Exemple 3 – Création d'un diagramme à tiges et à feuilles ordonné

On a demandé à 15 personnes d'indiquer combien de fois elles se sont rendues au travail au volant d'un véhicule au cours d'une période de 10 jours ouvrables.Voici le nombre de fois déclaré par chaque personne :

5, 7, 9, 9, 3, 5, 1, 0, 0, 4, 3, 7, 2, 9, 8

Créez un diagramme à tiges et à feuilles ordonné pour ce tableau.

On devrait dessiner ce tracé comme suit :

Tableau 4. Nombre de fois où chaque personne s'est rendue au travail au volant d'un véhicule pendant 10 jours

Tige	Feuille
0	0 0 1 2 3 3 4 5 5 7 7 8 9 9 9

Division des tiges

L'organisation de ce diagramme à tiges et à feuilles ne fournit pas beaucoup d'information sur les données. Le fait de n'avoir qu'une seule tige crée une feuille surchargée. Si les feuilles sont trop chargées, il pourrait alors être utile de diviser chaque tige en deux composantes ou plus. Ainsi, on peut diviser un intervalle 0 à 9 en deux intervalles de 0 à 4 et de 5 à 9. De même, on pourrait diviser une tige 0 à 9 en cinq intervalles : 0-1, 2-3, 4-5, 6-7 et 8-9.

Voici ce à quoi le diagramme à tiges et à feuilles devrait alors ressembler :

Tableau 5. Nombre de fois où chaque personne s'est rendue au travail au volant d'un véhicule pendant 10 jours

Tige	Feuille
$0^{(0)}$	0 0 1 2 3 3 4
$0^{(5)}$	5 5 7 7 8 9 9 9

Nota : La tige $0^{(0)}$ signifie que toutes les données se situent à l'intérieur de l'intervalle 0 à 4. La tige $0^{(5)}$ signifie que toutes les données se situent à l'intérieur de l'intervalle 5 à 9.

Exemple 4 – Division des tiges

Gabrielle est une nageuse qui s'entraîne pour une compétition. Voici le nombre de longueurs de 50 mètres qu'elle a parcourues à la nage quotidiennement pendant 30 jours :

22, 21, 24, 19, 27, 28, 24, 25, 29, 28, 26, 31, 28, 27, 22, 39, 20, 10, 26, 24, 27, 28, 26, 28, 18, 32, 29, 25, 31, 27

1. Tracez un diagramme à tiges et à feuilles ordonné. Formulez un bref commentaire sur ce qu'il indique.
2. Redessinez le diagramme à tiges et à feuilles en divisant les tiges en intervalles de cinq unités. Formulez un bref commentaire sur ce qu'indique le nouveau tracé.

Réponses

1. La valeur des observations varie de 10 à 39; le diagramme à tiges et à feuilles devrait donc comporter des tiges de 1, 2 et 3. Le diagramme à tiges et à feuilles ordonné est illustré ci-dessous :

   Tableau 6. Longueurs que Gabrielle a parcourues à la nage en 30 jours

Tige	Feuille
1	0 8 9
2	0 1 2 2 4 4 4 5 5 6 6 6 7 7 7 7 8 8 8 8 8 9 9
3	1 1 2 9

   Le diagramme à tiges et à feuilles montre que Gabrielle parcourt habituellement à la nage de 20 à 29 longueurs pendant sa séance quotidienne d'entraînement.

2. La division des tiges en intervalles de cinq unités donne le diagramme à tiges et à feuilles suivant :

   Tableau 7. Longueurs que Gabrielle a parcourues à la nage en 30 jours

Tige	Feuille
$1^{(0)}$	0
$1^{(5)}$	8 9
$2^{(0)}$	0 1 2 2 4 4 4
$2^{(5)}$	5 5 6 6 6 7 7 7 7 8 8 8 8 8 9 9
$3^{(0)}$	1 1 2
$3^{(5)}$	9

   Nota : La tige $1^{(0)}$ signifie que toutes les données se situent entre 10 et 14, alors que $1^{(5)}$ signifie que toutes les données se situent entre 15 et 19, et ainsi de suite.

   Le diagramme à tiges et à feuilles révisé montre que Gabrielle parcourt habituellement de 25 à 29 longueurs durant sa séance quotidienne d'entraînement. Les valeurs $1^{(0)}$ 0 = 10 et $3^{(5)}$ 9 = 39 sont des valeursaberrantes, un concept qui sera décrit dans la section suivante.

Exemple 5 – Division des tiges à l'aide de valeurs décimales

Voici les mesures de poids de 30 élèves (arrondis au dixième de kilogramme) qui ont été enregistrés :

59,2, 61,5, 62,3, 61,4, 60,9, 59,8, 60,5, 59,0, 61,1, 60,7, 61,6, 56,3, 61,9, 65,7, 60,4, 58,9, 59,0, 61,2, 62,1, 61,4, 58,4, 60,8, 60,2, 62,7, 60,0, 59,3, 61,9, 61,7, 58,4, 62,2

Tracez un diagramme à tiges et à feuilles ordonné pour les données. Formulez un bref commentaire sur ce qu'indique l'analyse.

Réponse

Dans ce cas, les tiges seront les valeurs exprimées en nombres entiers et les feuilles, les valeurs décimales. Les données varient de 56,3 à 65,7; les tiges devraient donc commencer à 56 et se terminer à 65.

Tableau 8. Poids de 30 élèves

Tige	Feuille
56	3

57	
58	4 4 9
59	0 0 2 3 8
60	0 2 4 5 7 8 9
61	1 2 4 4 5 6 7 9 9
62	1 2 3 7
63	
64	
65	7

Dans cet exemple, il n'était pas nécessaire de diviser les tiges, parce que les feuilles ne sont pas entassées sur un trop petit nombre de tiges, ni d'arrondir les valeurs, puisque leur plage n'est pas trop étendue. Ce diagramme à tiges et à feuilles révèle que le groupe pour lequel on a enregistré le nombre le plus élevé d'observations est celui de 61,0 à 61,9.

Valeurs aberrantes

Une *valeur aberrante* est une valeur extrême des données. C'est une valeur d'observation énormément différente du reste de ces dernières. Il peut y avoir plus d'une valeur aberrante dans un ensemble de données.

Dans l'exemple précédent, 56,3 et 65,7 sont des valeurs aberrantes, puisque ces deux valeurs sont assez différentes des autres.

Si l'on ne tient pas compte de ces deux valeurs aberrantes, on pourrait redessiner le diagramme à tiges et à feuilles de l'exemple précédent comme suit :

Tableau 9. Poids de 30 élèves, à l'exception des valeurs aberrantes

Tige	Feuille
58	4 4 9
59	0 0 2 3 8
60	0 2 4 5 7 8 9
61	1 2 4 4 5 6 7 9 9
62	1 2 3 7

Lorsqu'on utilise un diagramme à tiges et à feuilles, repérer une valeur aberrante est souvent une question de jugement, parce qu'il n'existe pas de règle stricte à respecter relativement à la distance devant séparer une valeur du reste d'un ensemble de données pour qu'elle remplisse les conditions nécessaires afin d'être considérée comme une valeur aberrante, sauf dans le cas où l'on emploierait des tracés en rectangle (tel qu'il est décrit dans la section sur les diagrammesdequartiles).

Caractéristiques des distributions

Lorsque vous évaluez le profil général d'une distribution (c'est-à-dire le profil établi à partir de toutes les valeurs d'une variable particulière), recherchez ces caractéristiques :

- le nombre de pointes de la distribution;
- sa forme générale (désaxée [asymétrique] ou symétrique);
- son centre;
- sa dispersion.

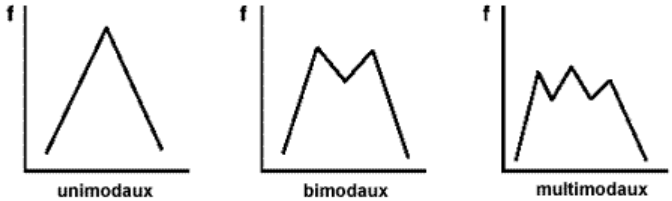
Nombre de pointes

Les diagrammeslinéaires sont utiles parce qu'ils révèlent facilement et rapidement certaines caractéristiques des données.

La première caractéristique qu'on peut facilement et rapidement voir à partir d'un diagramme linéaire est le nombre de points élevés ou de pointes élevées que compte la distribution.

Même si la plupart des distributions de données statistiques ne comptent qu'une pointe principale (distributions *unimodales*), d'autres peuvent compter deux pointes (distributions *bimodales*) ou plus de deux pointes (distributions *multimodales*).

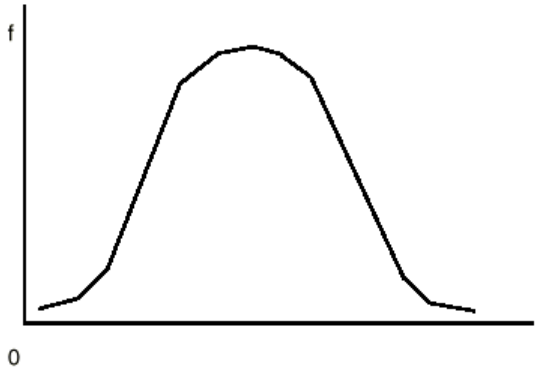
Voici des exemples de diagrammes linéaires unimodaux, bimodaux et multimodaux :



Forme générale

La deuxième caractéristique principale d'une distribution est son degré de *symétrie*.

Une courbe *parfaitement symétrique* est une courbe dont les deux côtés de la distribution se marieraient exactement l'un à l'autre si l'on pliait la figure en son point central. En voici un exemple :

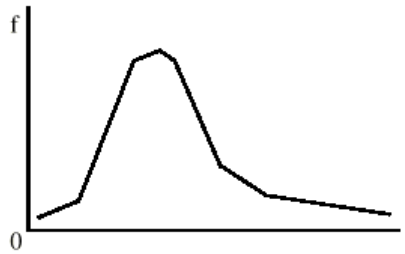


On appelle une distribution symétrique, unimodale, en (forme de) cloche, un cas relativement courant, une distribution normale.

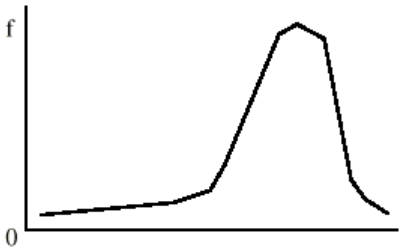
Si la distribution est dissymétrique, on dit qu'elle est *désaxée*.

On dit qu'une distribution est désaxée vers la droite, ou *positivement désaxée*, lorsque la plupart des données sont concentrées à sa gauche. Les distributions positivement désaxées sont plus courantes que les distributions négativement désaxées.

Le revenu fournit un exemple d'une distribution positivement désaxée. La plupart des gens gagnent moins de 40 000 $ par année, mais certains gagnent un revenu nettement plus élevé et un plus petit nombre gagne plusieurs millions annuellement. La queue positive (droite) du diagramme linéaire pour le revenu s'étend donc assez loin, tandis que la queue de l'asymétrie négative (gauche) s'arrête à zéro. La queue droite s'étend manifestement plus loin à partir du centre de la distribution que la queue gauche, comme il est indiqué ci-dessous :



On dit qu'une distribution est désaxée vers la gauche, ou *négativement désaxée*, si la plupart des données sont concentrées à sa droite. La queue gauche s'étend manifestement plus loin à partir du centre de la distribution que la queue droite, comme il est indiqué ci-dessous :



Centre et dispersion

On peut situer le centre (la *médiane*) d'une distribution en comptant la moitié des observations à partir de la plus petite. Évidemment, cette méthode est impossible à utiliser pour de très grands ensembles de données. Un diagramme à tiges et à feuilles en facilite cependant l'utilisation, parce que les données sont disposées dans l'ordre ascendant. (Voir le chapitre sur les mesuresdetendancecentrale pour plus de détails.)

On peut rapidement repérer sur un diagramme l'ampleur de la dispersion d'une distribution et tous les écarts importants par rapport au profil général (les valeurs aberrantes).

Utilisation de diagrammes à tiges et à feuilles comme diagrammes

Un diagramme à tiges et à feuilles est un type simple de diagramme qui est dessiné à partir des nombres eux-mêmes. C'est un moyen de présenter les principales caractéristiques d'une distribution. Si on le tourne de côté, un diagramme à tiges et à feuilles ressemblera à un histogramme ou à un diagramme à bandes et fournira de l'information visuelle similaire.

Exemple 6 – Utilisation de diagrammes à tiges et à feuilles comme diagrammes

Voici les résultats enregistrés d'examens de mathématiques de 41 élèves (pour lesquels la meilleure note possible était de 70) :

31, 49, 19, 62, 50, 24, 45, 23, 51, 32, 48, 55, 60, 40, 35, 54, 26, 57, 37, 43, 65, 50, 55, 18, 53, 41, 50, 34, 67, 56, 44, 4, 54, 57, 39, 52, 45, 35, 51, 63, 42

1. La variable est-elle discrète ou continue? Expliquez.
2. Tracez un diagramme à tiges et à feuilles ordonné pour les données et décrivez brièvement ce qu'il indique.
3. Y a-t-il des valeurs aberrantes? Si oui, quelles sont ces notes?
4. Examinez de côté le diagramme à tiges et à feuilles. Décrivez les principales caractéristiques de la distribution comme :
    a. le nombre de pointes;

b. la symétrie;

c. la valeur au centre de la distribution.

Réponses

1. Une note d'examen est une variable discrète. Il est impossible, par exemple, d'avoir une note d'examen de 35,74542341...
2. La valeur la plus faible est 4 et la valeur la plus élevée, 67. Voici donc à quoi ressemble le diagramme à tiges et à feuilles qui englobe cette plage de valeurs.

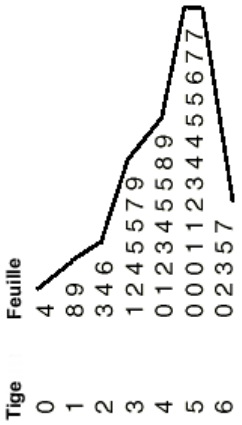
Tableau 10. Notes d'examens de mathématiques de 41 élèves

Tige	Feuille
0	4
1	8 9
2	3 4 6
3	1 2 4 5 5 7 9
4	0 1 2 3 4 5 5 8 9
5	0 0 0 1 1 2 3 4 4 5 5 6 7 7
6	0 2 3 5 7

Nota : La notation 2|4 représente la *tige* 2 et la *feuille* 4.

Le diagramme à tiges et à feuilles révèle que la plupart des élèves ont obtenu des notes qui se situent à l'intérieur de l'intervalle de 50 à 59. Le grand nombre d'élèves qui ont obtenu des résultats élevés pourrait indiquer que l'examen était trop facile, que la plupart des élèves connaissaient bien la matière, ou une combinaison des deux.

3. Le résultat de 4 pourrait être une valeur aberrante, puisqu'il y a un large écart entre ce résultat et le résultat suivant, 18.
4. Si on le tourne de côté, le diagramme à tiges et à feuilles ressemblera à ce qui suit :



La distribution ne comporte qu'une seule pointe, qui se situe à l'intérieur de l'intervalle 50 à 59.

Même s'il n'y a que 41 observations, la distribution montre que la plupart des données sont groupées à droite. La queue gauche s'étend plus loin du centre des données que la queue droite. La distribution est donc désaxée vers la gauche ou *négativement désaxée*.

Puisqu'il y a 41 observations, le centre de la distribution se situera à la 21e observation. Si l'on compte 21 observations à partir de la plus petite, le centre est 48. (Notez que l'on aurait obtenu la même valeur si l'on avait compté 21 observations à partir de la plus élevée.)

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. le nombre d'employés à Statistique Canada;
    j. le nombre de frères et sœurs que vous avez;
    k. la distance entre votre domicile et votre école;
    l. le nombre de pages dans un dictionnaire.

2. Sans utiliser les exemples tirés de la question 1, donnez deux autres exemples :
    a. d'une variable discrète
    b. d'une variable continue

3. Une compagnie de téléphone a fait enquête auprès de 12 ménages pour déterminer combien d'appareils téléphoniques il y avait par ménage.
    a. Copiez dans votre cahier le tableau de distribution de fréquences ci-dessous et remplissez-le à l'aide des résultats d'enquête qui suivent :

    2, 5, 4, 3, 4, 3, 1, 3, 3, 2, 3, 4

Question 3a

Nombre d'appareils téléphoniques (x)	Comptage	Fréquence (f)
1		
2		
3		
4		
5		

    b. Quel est le résultat le plus fréquent?

4. 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. S'agit-il de variables discrètes ou continues?
    b. Présentez ces données au moyen d'un tableau de distribution de fréquences.
    c. Quel est le résultat le plus fréquent?
    d. Créez un tableau de distribution de fréquences qui inclue des colonnes pour la fréquence relative et la fréquence en pourcentage des données.
    e. Quelles conclusions pouvez-vous tirer des tableaux? Expliquez.

5. Un vent a soufflé pendant 40 jours. Voici les vitesses enregistrées :

    15, 22, 14, 12, 21, 34, 19, 11, 13, 0, 16, 4, 23, 8, 12, 18, 24, 17, 14, 3, 10, 12, 9, 15, 20, 5, 19, 13, 17, 11, 16, 19, 24, 12, 7, 14, 17, 10, 14, 23

a. S'agit-il de variables discrètes ou continues?

b. Choisissez un intervalle de classe approprié et présentez ces données au moyen d'un tableau de distribution de fréquences.

c. Quel est l'intervalle de classe le plus fréquent?

d. Créez un tableau de distribution de fréquences qui inclue des colonnes pour la fréquence relative et la fréquence en pourcentage des données.

e. Quelles conclusions peut-on tirer des tableaux? Expliquez.

6.

a. Tracez un diagramme à tiges et à feuilles ordonné pour les données incluses dans la question 5.

b. Y a-t-il des valeurs aberrantes? Si oui, pourquoi?

c. Décrivez les principales caractéristiques de la distribution :

i. le nombre de pointes;

ii. la forme générale;

iii. la valeur approximative au centre de la distribution.

7. On a fait enquête auprès de 30 personnes pour déterminer à quelle fréquence elles allaient au cinéma dans une année. Voici les résultats de cette enquête :

21, 35, 27, 2, 18, 25, 10, 4, 43, 14, 29, 24, 15, 9, 26, 31, 41, 1, 28, 38, 40, 22, 37, 26, 19, 0, 33, 12, 16, 23

a. Copiez dans votre cahier le diagramme à tiges et à feuilles figurant ci-dessous et entrez-y les résultats.

Question 7a

Tige	Feuille
0	
1	
2	
3	
4	

b. Maintenant, convertissez-le en un diagramme à tiges et à feuilles ordonné.

8. Supposez les nombres annuels suivant d'accidents mortels survenus sur les routes de 1960 à 1992 :

10, 7, 8, 8, 17, 15, 17, 23, 14, 26, 31, 20, 32, 29, 31, 32, 38, 29, 30, 24, 30, 29, 26, 28, 37, 33, 32, 36, 32, 32, 26, 17, 20

a. S'agit-il de variables discrètes ou continues?

b. Tracez un diagramme à tiges et à feuilles de ces données.

c. Agrandissez le diagramme à l'aide d'intervalles de classe de cinq unités.

d. Y a-t-il des valeurs aberrantes? Si oui, pourquoi?

e. Décrivez les principales caractéristiques de la distribution :

i. le nombre de pointes;

ii. la forme générale;

iii. la valeur approximative au centre de la distribution.

9. Voici la température moyenne minimale en avril (en degrés Celsius ou ºC) enregistrée de 1982 à 2002 :

6,1, 8,9, 6,9, 7,2, 7,0, 6,2, 5,7, 6,2, 6,8, 6,4, 6,8, 6,4, 7,6, 7,8, 7,3, 6,8, 8,8, 7,8, 8,1, 8,1, 7,9

a. S'agit-il de variables discrètes ou continues?

b. Tracez un diagramme à tiges et à feuilles ordonné de ces données.

c. Faut-il agrandir le diagramme à tiges et à feuilles? Si oui, pourquoi? Si non, pourquoi?

d. Y a-t-il des valeurs aberrantes? Si oui, pourquoi?

e. Décrivez les principales caractéristiques de la distribution.

i. le nombre de pointes;

ii. la forme générale;

iii. la valeur approximative au centre de la distribution.

10. On a fait enquête auprès de 50 membres du personnel d'une entreprise de construction pour déterminer quel était leur salaire hebdomadaire arrondi au dollar près. Voici les résultats de cette enquête :

514, 476, 497, 511, 484, 513, 471, 470, 441, 466, 443, 481, 502, 528, 459, 548, 521, 517, 463, 478, 473, 514, 542, 519, 522, 523, 546, 487, 486, 473, 527, 470, 440, 564, 499, 523, 484, 463, 461, 437, 555, 525, 461, 539, 466, 470, 486, 490, 543, 519

a. S'agit-il de variables discrètes ou continues?

b. Choisissez un intervalle de classe approprié et présentez ces données au moyen d'un tableau de distribution de fréquences.

c. Quel est l'intervalle de classe le plus fréquent?

d. Créez un tableau de distribution de fréquences qui inclue des colonnes pour la fréquence relative et la fréquence en pourcentage des données.

e. Quelles conclusions pouvez-vous tirer des tableaux? Expliquez.

f. Tracez un diagramme à tiges et à feuilles ordonné de ces données.

g. Y a-t-il des valeurs aberrantes? Si oui, pourquoi?

h. Décrivez les principales caractéristiques de la distribution :

i. le nombre de pointes;

ii. la forme générale;

iii. la valeur approximative au centre de la distribution.

Activités en classe

1. Tracez une ligne droite de 10 centimètres (cm) de longueur exactement. Sans mesurer, faites une coche là où vous estimez que se situe le point milieu. Mesurez maintenant la ligne et placez une coche au point milieu réel (à 5 cm). Mesurez la distance entre votre estimation et le point milieu réel. À combien de millimètres (mm) votre estimation se trouvait-elle par rapport au point milieu réel?

   a. Inscrivez cette valeur dans un tableau. Déterminez à quelle distance chacun des autres élèves de la classe s'écartait du point milieu et inscrivez ces résultats.

   b. Avec ces données, créez un tableau de distribution de fréquences qui inclue des colonnes pour la fréquence relative et la fréquence en pourcentage des données.

   c. Quel est le résultat le plus fréquent?

   d. Tracez un diagramme à tiges et à feuilles de ces données.

   e. Y a-t-il des erreurs aberrantes? Si oui, pourquoi?

   f. Décrivez les principales caractéristiques de la distribution :

      i. le nombre de pointes

      ii. la forme générale

      iii. la valeur approximative au centre de la distribution

   g. Quelles conclusions pouvez-vous tirer de cette analyse?

2. Demandez à votre enseignante ou à votre enseignant les notes (anonymes) d'un récent test ou travail de la classe. Effectuez une analyse détaillée de ces données à l'aide des instructions fournies aux points a) à g) de la question 1. Ensuite, formulez un bref commentaire sur :

   a. la moyenne de la classe pour le test ou le travail;

   b. la capacité de la classe de comprendre les questions du test ou du travail;

   c. l'intérêt que la classe semble avoir pour la matière ayant fait l'objet du test ou du travail.

   Fournissez pour chaque commentaire des preuves à l'appui tirées de votre analyse.

3. Lancez un dé 30 fois. Inscrivez chaque fois le résultat dans un tableau de distribution de fréquences.

   a. S'agit-il de variables discrètes ou continues?

   b. Créez un tableau de distribution de fréquences qui inclue des colonnes pour la fréquence relative et la fréquence en pourcentage des données.

   c. Quel est le résultat le plus fréquent?

   d. Vous attendiez-vous à ce qu'un nombre apparaisse plus souvent que les autres? Si oui, pourquoi?

   e. Tracez un diagramme à tiges et à feuilles de ces données.

   f. Y a-t-il des valeurs aberrantes? Si oui, pourquoi?

   g. Décrivez les principales caractéristiques de la distribution:

      i. le nombre de pointes;

      ii. la forme générale;

      iii. la valeur approximative au centre de la distribution.

   h. Quelles conclusions pouvez-vous tirer de l'analyse?

4. Créez un tableau indiquant des couleurs primaires et secondaires : le rouge, le bleu, le jaune, le violet, le vert et l'orange. Incluez également le noir, le blanc et le gris. N'incluez pas dans votre tableau de teintes ou de tons complémentaires (le bleu-vert, le mauve, le beige, etc.). Enfin, intitulez la dernière colonne « Aucune de ces couleurs ».

   Faites enquête auprès des enseignants de votre école pour déterminer la couleur de leur véhicule. Si un enseignant mentionne une couleur qui ne figure pas dans votre tableau, inscrivez sa réponse dans la colonne « Aucune de ces couleurs ».

   a. S'agit-il de variables discrètes ou continues?

b. Créez un tableau de distribution de fréquences qui inclue des colonnes pour la fréquence relative et la fréquence en pourcentage des données.

c. Déterminez la couleur de véhicule la plus populaire chez les enseignants questionnés. Dans quelle proportion cette couleur est-elle plus populaire que la deuxième couleur la plus courante?

d. Pourquoi est-il impossible de dessiner un diagramme à tiges et à feuilles de ces données?

e. Pour quelle raison un manufacturier de véhicules automobiles serait-il intéressé par ce type d'analyse de données?

Réponses

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. Il s'agit d'une variable continue.

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 discrète.

j. Il s'agit d'une variable discrète.

k. Il s'agit d'une variable continue.

l. Il s'agit d'une variable discrète.

2. Diverses réponses.

3.

a.

Réponses, question 3a

Nombre d'appareils téléphoniques (x)	Comptage	Fréquence (f)
1	I	1
2	II	2
3	𝍸	5
4	III	3
5	I	1
Total		**12**

b. Le résultat le plus fréquent est 3.

4.

a. Il s'agit de variables discrètes.

b.

Réponses, question 4b

Nombre de clients (x)	Comptage	Fréquence (f)
20	II	2
21	𝍸 II	7
22	IIII	4
23	III	3
24	𝍸	5
25	II	2
26	II	2
Total		**25**

c. L'observation la plus fréquente est 21.

d.

Réponses, question 4d

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**	**100,0**

5.

a. Ce sont des variables continues.

b.

Réponses, question 5b

Vitesse du vent (x)	Comptage	Fréquence (f)
0 à < 5	III	3
5 à < 10	IIII	4
10 à < 15	𝍸 𝍸 IIII	14
15 à < 20	𝍸 𝍸 I	11
20 à < 25	𝍸 II	7
25 à < 30		0
30 à < 35	I	1
Total		**40**

c. L'intervalle de classe qui apparaît le plus souvent est 10 à < 15

d.

Réponses, question 5d

Vitesse du vent (x)	Fréquence (f)	Fréquence relative	Fréquence en pourcentage
0 à < 5	3	0,075	7,5
5 à < 10	4	0,100	10,0
10 à < 15	14	0,350	35,0
15 à < 20	11	0,275	27,5
20 à < 25	7	0,175	17,5
25 à < 30	0	0,000	0,0
30 à < 35	1	0,025	2,5
Total	**40**	**1,00**	**100,0**

e. La vitesse du vent la plus fréquente varie de 10 à moins de 15 nœuds. D'après cet échantillon de 40 jours, la possibilité que le vent atteigne cette vitesse n'importe quelle journée est de 35 %.

6.

a.

Réponses, question 6a

Tige	Feuille
$0^{(0)}$	0 3 4
$0^{(5)}$	5 7 8 9
$1^{(0)}$	0 0 1 1 2 2 2 2 3 3 4 4 4 4
$1^{(5)}$	5 5 6 6 7 7 7 8 9 9 9
$2^{(0)}$	0 1 2 3 3 4 4

$2^{(5)}$	
$3^{(0)}$	4

b. La valeur aberrante de cet exercice est 34. Explications possibles :

- il y a eu durant les 40 jours une journée particulièrement venteuse ou orageuse;
- il y avait une erreur de mesure.

c.

i. La distribution ne compte qu'une seule pointe principale.

ii. La distribution est à peu près asymétrique. Si l'on soustrayait l'erreur aberrante, la distribution serait un peu désaxée vers la gauche. La distribution revêt généralement une forme irrégulière.

iii. Le centre de la distribution est 14 nœuds.

7.

a.

Réponses, question 7a

Tige	Feuille
0	2 4 9 1 0
1	8 0 4 5 9 2 6
2	1 7 5 9 4 6 8 2 6 3
3	5 1 8 7 3
4	3 1 0

b.

Réponses, question 7b

Tige	Feuille
0	0 1 2 4 9
1	0 2 4 5 6 8 9
2	1 2 3 4 5 6 6 7 8 9
3	1 3 5 7 8
4	0 1 3

8.

a. Ce sont des variables discrètes.

b.

Réponses, question 8b

Tige	Feuille
0	7 8 8
1	0 4 5 7 7 7
2	0 0 3 4 6 6 6 8 9 9 9
3	0 0 1 1 2 2 2 2 2 3 6 7 8

c.

Réponses, question 8c

Tige	Feuille
$0^{(5)}$	7 8 8
$1^{(0)}$	0 4

$1^{(5)}$	5 7 7 7
$2^{(0)}$	0 0 3 4
$2^{(5)}$	6 6 6 8 9 9 9
$3^{(0)}$	0 0 1 1 2 2 2 2 2 3
$3^{(5)}$	6 7 8

d. Il n'y a pas de valeur aberrante dans cet exercice.

e.

i. La distribution ne compte qu'une seule pointe principale.

ii. La forme générale de la distribution est désaxée vers la gauche.

iii. La valeur approximative au centre de la distribution est 28 accidents mortels sur les routes ontariennes.

9.

a. Ce sont des variables continues.

b.

Réponses, question 9b

Tige	Feuille
5	7
6	1 2 2 4 4 8 8 8 9
7	0 2 3 6 8 8 9
7	1 1 8 9

c. Non, il n'est pas nécessaire d'agrandir le diagramme parce que les tiges ne sont pas surchargées.

d. Il n'y a aucune valeur aberrante.

e.

i. La distribution ne compte qu'une seule pointe principale.

ii. La forme générale de la distribution est à peu près symétrique (même s'il est difficile de l'observer avec un faible nombre de données).

iii. Le centre de la distribution est 7°C.

10.

a. Ce sont des variables discrètes.

b.

Réponses, question 10b

Salaire hebdomadaire	Comptage	Fréquence (f)
420 à < 440	I	1
440 à < 460	IIII	4
460 à < 480	𝍸 𝍸 IIII	14
480 à < 500	𝍸 IIII	9
500 à < 520	𝍸 III	8
520 à < 540	𝍸 III	8
540 à < 560	𝍸	5
560 à < 580	I	1
Total		**50**

c. L'intervalle de classe le plus fréquent est 460 $ à < 480 $.

d.

Réponses, question 10d

Salaire hebdomadaire (x)	Fréquence (f)	Fréquence relative	Fréquence en pourcentage
420 à < 440	1	0,02	2
440 à < 460	4	0,08	8
460 à < 480	14	0,28	28
480 à < 500	9	0,18	18
500 à < 520	8	0,16	16
520 à < 540	8	0,16	16
540 à < 560	5	0,10	10
560 à < 580	1	0,02	2
Total	**50**	**1,00**	**100**

e. Le groupe salarial le plus courant, qui représente 28 % des employés de l'entreprise, se situe entre 460 $ et 480 $ par semaine d'après cet échantillon de 50 employés. Un seul employé gagnait plus de 560 $ par semaine.

f.

Réponses, question 10f

Tige	Feuille
43	7
44	0 1 3
45	9
46	1 1 3 3 6 6
47	0 0 0 1 3 3 6 8
48	1 4 4 6 6 7
49	0 7 9
50	2
51	1 3 4 4 7 9 9
52	2 3 3 5 7 8
53	9
54	2 3 6 8
55	5
56	4

g. Il n'y a aucune valeur aberrante. S'il y en avait eu, voici quelles auraient pu être les explications possibles :

- les employés qui gagnaient ces salaires étaient peut-être des cadres ou des directeurs et, par conséquent, mieux rémunérés;
- ces employés ont peut-être fourni des réponses trompeuses.

h.

i. La distribution compte plusieurs pointes, ce qui indique la possibilité qu'elle soit bimodale.

ii. La distribution n'est ni symétrique ni désaxée.

iii. Le centre de la distribution se situe entre 487 $ et 490 $.

Types de diagrammes

Les 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 diagramme peut montrer une tendance ou faire une comparaison.

Les élèves considèrent habituellement les diagrammes faciles à utiliser parce qu'ils sont constitués de lignes, de points et de blocs – des formes géométriques simples qu'ils peuvent rapidement dessiner.

Dans le monde des statistiques, les diagrammes montrent les relations entre des variables ou l'étendue des valeurs d'une variable ou d'un phénomène donné.

Utilisation des diagrammes

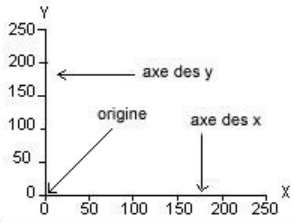
Qu'est-ce qu'un diagramme?

Un diagramme est, entre autres, une représentation visuelle d'une relation entre deux variables. Un diagramme est habituellement une figure en une ou deux dimensions comme un nuage de points. Même s'il existe des diagrammes en trois dimensions, ils sont habituellement considérés trop complexes pour être compris facilement.

Un diagramme est habituellement formé de deux axes appelés axe des x (horizontal) et axe des y (vertical). Ces axes correspondent aux variables utilisées comme information. Les axes ont un nom différent, comme *Prix* et *Quantité*.

Le point de rencontre des deux axes est appelé origine. L'origine correspond également au point (0,0).

Figure 1. Parties d'un diagramme



Un point dans un diagramme représente une relation. Chaque point est défini par deux coordonnées (x et y). Une coordonnée est un ensemble de nombres indiquant la position d'un point dans le diagramme.

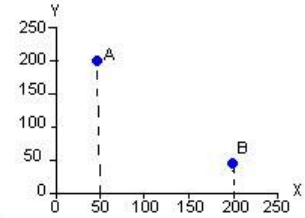
Dans la section suivante, vous apprendrez comment déterminer les coordonnées d'un point et de bien noter les coordonnées de ce point.

Identification de la coordonnée de l'axe des x

La coordonnée de l'axe des x d'un point, ou abscisse, correspond à la valeur indiquant la distance entre le point et l'origine (horizontale). Pour trouver la coordonnée de l'axe des x d'un point dans un diagramme, tracez, à partir du point, une ligne perpendiculaire à l'axe des x. Le nombre correspondant à la rencontre de la ligne avec l'axe des x est la valeur sur l'axe des x.

La figure 2 comporte un diagramme où se trouvent deux points, A et B. Identifiez la coordonnée sur l'axe des x des points A et B.

Figure 2. Coordonnée sur l'axe des x



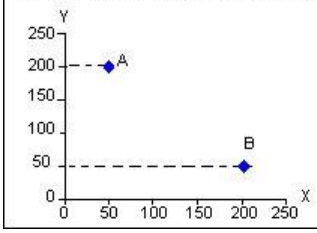
Réponse : La coordonnée du point A sur l'axe des x est 50 et la coordonnée du point B sur l'axe des x est 200.

Identification de la coordonnée sur l'axe des Y

La coordonnée d'un point sur l'axe des y, ou ordonnée, est la valeur indiquant la distance entre le point et l'origine sur l'axe vertical ou axe des y. Pour trouver la coordonnée de l'axe des y d'un point dans un diagramme, tracez, à partir du point, une ligne perpendiculaire à l'axe des y. Le nombre correspondant à la rencontre de la ligne avec l'axe des y est la valeur sur l'axe des y.

Identifiez la coordonnée sur l'axe des y des points A et B de la figure 3.

Figure 3. Coordonnée sur l'axe des y

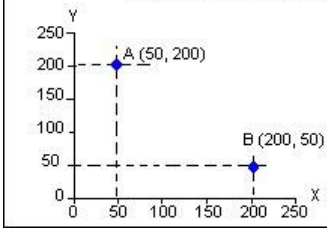


Réponse : La coordonnée du point A sur l'axe des y est 200 et celle du point B est 50.

Identification des points dans un diagramme

Une fois que vous avez déterminé les coordonnées d'un point, vous pouvez identifier les points à l'aide de la notation par paires. Cette forme de notation est simple – les points sont identifiés en inscrivant leurs coordonnées de la façon suivante : (x, y). Il est important de noter qu'il faut d'abord inscrire la coordonnée de l'axe des x, tel qu'il est indiqué à la figure 2. Les coordonnées de l'axe des x et de l'axe des y pour les points A et B sont données à la figure 4 ci-dessous.

Figure 4. Identification des coordonnées à l'aide de la notation par paires

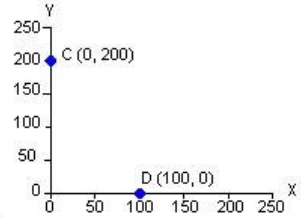


- L'abscisse du point A est 50 et son ordonnée est 200. Les coordonnées du point A sont donc (50, 200).
- L'abscisse du point B est 200 et son ordonnée est 50. Les coordonnées du point B sont donc (200, 50).

Points sur les axes

Si un point se situe sur un axe, il est inutile de tracer une ligne pour trouver les coordonnées. Dans la figure 5 ci-dessous, le point C se situe sur l'axe des y et le point D se situe sur l'axe des x. Lorsqu'un point se trouve directement sur un axe, une des coordonnées doit nécessairement être 0.

Figure 5. Points sur les axes



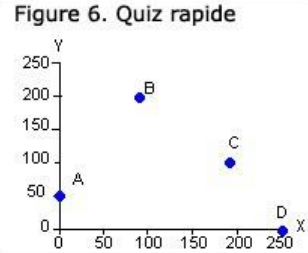
- Le point C se trouve sur l'axe des y et son abscisse est 0. En calculant la distance sur l'axe des y, vous verrez que le point est à 200 de l'origine. Les coordonnées du point C sont donc (0, 200).
- Le point D se trouve sur l'axe des x et son ordonnée est 0. En calculant la distance sur l'axe des x, vous verrez que le point est à 100 de l'origine. Les coordonnées du point D sont donc (100, 0).

Quiz rapide!

Répondez aux questions en utilisant la figure 6 ci-dessous.

- Quel point se situe sur l'axe des y?
- Quel point correspond à la notation par paires (100, 200)?
- Quels points ont comme ordonnée 100?

Figure 6. Quiz rapide



Réponses : 1. Point A 2. Point B 3. Point C

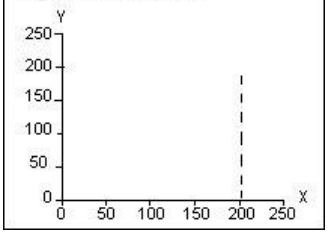
Placer des points dans un diagramme

Il pourra arriver qu'on vous donne les coordonnées d'un point et que vous ayez à en trouver le lieu sur un diagramme. C'est ce qu'on appelle placer un point dans un diagramme. Ci-dessous se trouvent des indications sur la façon de placer un point dans un diagramme.

Placez le point (200, 150) à l'aide de la méthode par étape donnée.

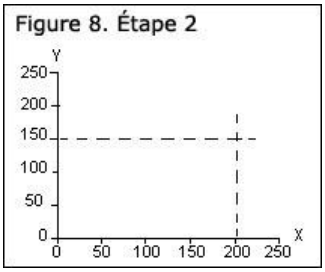
Étape 1
Premièrement, tracez une ligne perpendiculaire à partir de la coordonnée du point sur l'axe des x. Dans l'exemple, l'abscisse est 200.

Figure 7. Étape 1



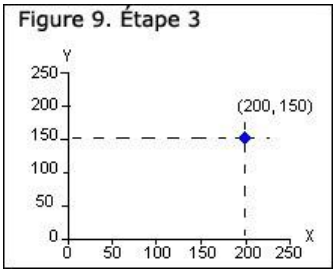
Étape 2
Tracez ensuite une ligne perpendiculaire à partir de la coordonnée du point sur l'axe des y. L'ordonnée dans l'exemple est 150.

Figure 8. Étape 2



Étape 3
Enfin, placez le point à l'intersection des deux lignes. C'est le point que nous voulons placer (200, 150).

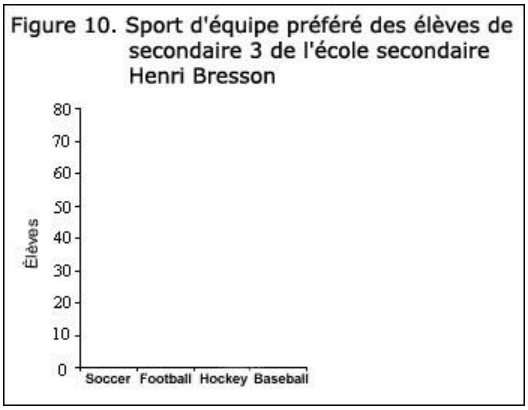
Figure 9. Étape 3



Choisir l'échelle

L'échelle d'un diagramme est très importante. Elle est déterminée par les données de chacun des axes.

Figure 10. Sport d'équipe préféré des élèves de secondaire 3 de l'école secondaire Henri Bresson



Un sondage a été effectué auprès des élèves de secondaire 3 à l'école secondaire Henri Bresson. Les élèves devaient indiquer lequel des sports d'équipe suivants ils préfèrent.

Les résultats étaient les suivants :

1. Soccer – 45 élèves
2. Football – 55 élèves
3. Hockey – 75 élèves
4. Baseball – 25 élèves

Dans la figure 10, ces quatre catégories ont été placées sur l'axe des x, chacune représentant les données recueillies. Parce qu'il s'agit de catégories nominales (des noms au lieu de chiffres) et qu'elles représentent un attribut qualitatif (non quantitatif), les groupes peuvent être placés dans n'importe quel ordre sur l'axe.

Sur l'axe des y, les données vont de 0 à 80 élèves. Comme mentionné plus haut, l'origine devrait être située à 0, là où les axes se rencontrent. Puisque le plus important groupe d'élèves pour un sport est 75, il est donc approprié que l'échelle s'arrête à 80. En changeant l'échelle, le diagramme ne changera pas, mais son aspect visuel sera modifié.

L'intervalle de l'échelle est l'espace entre les inscriptions sur l'axe. Si l'espace est trop petit, en général on le divise par 10. Faites-en votre intervalle. Pour les espaces plus grands, l'intervalle est habituellement 5, 10, 100, 500, 1 000, etc. Utilisez des nombres qui se divisent par 100, 1 000 ou leurs multiples afin que le diagramme soit facile à comprendre.

Dans le cas présent, si vous prenez 80 et que vous le divisez par 5, vous obtenez 16. Cependant, il serait peut-être mieux d'utiliser 10 pour faciliter l'analyse. L'échelle est plus petite, mais tout de même plus facile à utiliser.

Règles pour faire de bons diagrammes

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

Un bon diagramme :

- montre 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 diagramme pour présenter des données?

Les diagrammes :

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

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

- le diagramme à bandes
- le pictogramme
- le diagramme circulaire
- le diagramme linéaire
- le nuage de points
- l'histogramme

Il est primordial de connaître le bon type de diagramme à utiliser avec un type particulier d'information. Selon la nature des données, certains diagrammes sont plus appropriés que d'autres. Par exemple, il est préférable de présenter des données catégoriques (comme les sujets scolaires préférés) sous forme de diagramme à bandes ou de diagramme circulaire. Par ailleurs, les données numériques continues (comme la taille) s'illustrent mieux sous forme de diagramme linéaire ou d'histogramme. Pour plus de renseignements sur les types de diagrammes appropriés, voir « Types de données » dans le « Guide de l'enseignant sur la découverte des données ».

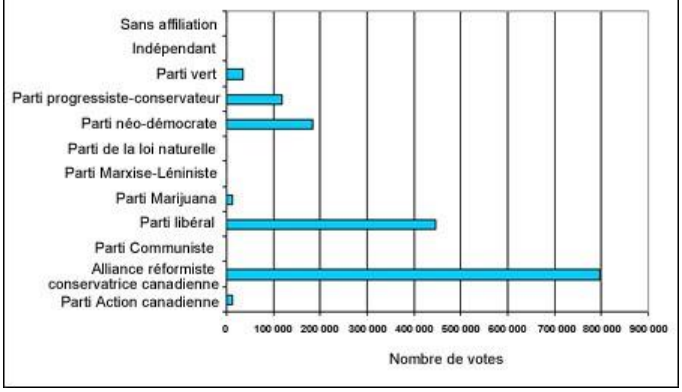
Quand n'est-il pas approprié d'utiliser un diagramme?

Le diagramme n'est pas toujours l'outil de choix pour présenter de l'information. Parfois, un texte ou un tableau de données fournit une meilleure explication — tout en vous ménageant de l'effort et du temps.

Pensez-y deux fois avant d'utiliser un diagramme dans les situations suivantes :

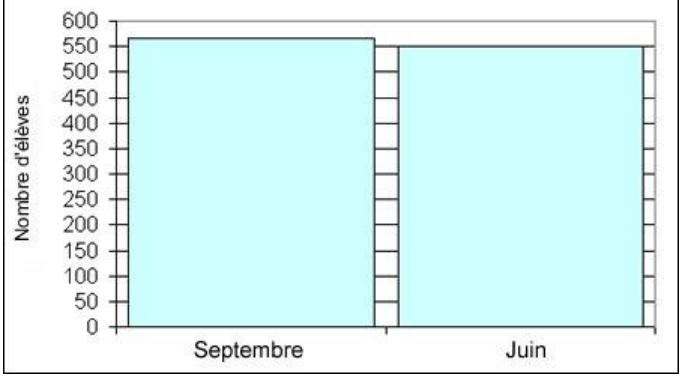
- les données sont trop dispersées

Figure 11. Répartition des votes pour les principaux partis politiques lors de l'élection fédérale, Uneville



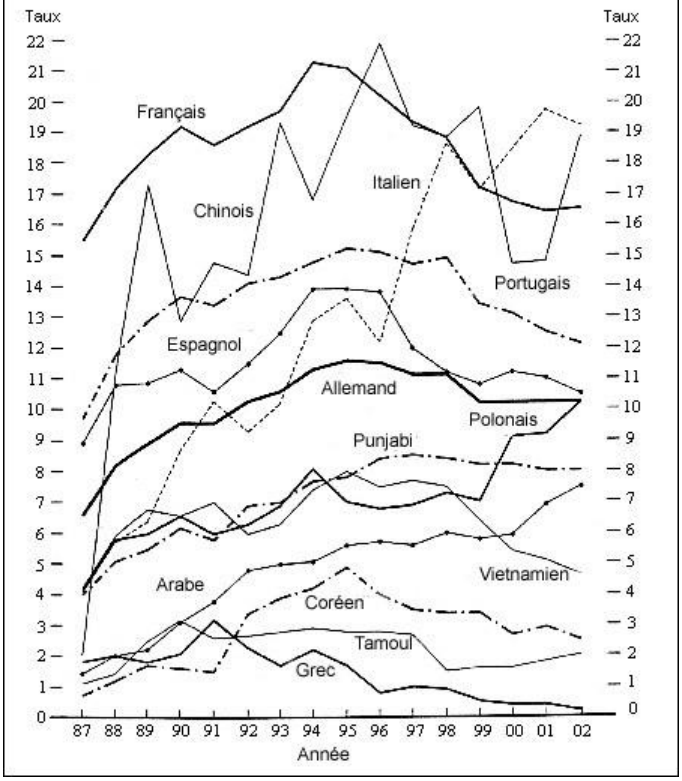
- il n'y a pas assez de données (seulement un, deux ou trois points de données)

Figure 12. Nombre d'élèves inscrits à l'école secondaire du Massif



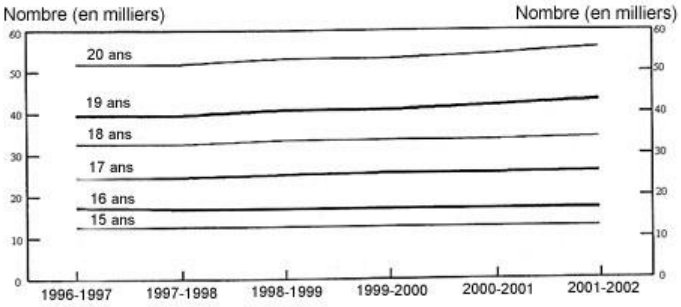
- les données sont trop nombreuses

Figure 13. Nombre d'élèves qui étudient l'anglais comme langue seconde à l'école secondaire des Pays d'en Haut, par langue maternelle, de 1987 à 2002



- les données ne sont pas très variées

Figure 14. Nombre de jeunes adultes qui font du sport au moins une fois par semaine, par âge, de 1996 à 2002



Diagrammes : quatre directives

Si vous avez conclu qu'un diagramme 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 :

   a. Qui est le public cible?
   b. Que connaît-il de la question?
   c. Que s'attend-il à voir?
   d. Que veut-il savoir?
   e. Que fera-t-il de l'information?

2. **Déterminez le message à être transmis.**

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

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

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

   Examinez le tableau suivant de termes appropriés à utiliser dans le titre d'un diagramme ou pour le décrire dans le texte accompagnateur :

Utilisez les termes appropriés pour décrire votre diagramme

Si votre diagramme...	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 données	augmente avec, diminue avec, varie avec, malgré, correspond à, est lié à

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

   a. diagramme circulaire (description des composantes)
   b. diagramme à bandes horizontales (comparaison des éléments et relations, série chronologique)
   c. diagramme à bandes verticales (comparaison des éléments et relations, série chronologique, distribution de fréquences)
   d. diagramme linéaire (série chronologique, distribution de fréquences)
   e. nuage de points (analyse des relations)

Diagramme à bandes

Un diagramme à bandes peut être horizontal ou vertical. Ce qu'il faut se rappeler au sujet des diagrammes à bandes est la longueur ou la hauteur des barres — plus elles sont longues ou hautes, plus leur valeur est grande.

Les diagrammes à bandes 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 diagrammes à bandes présentent habituellement des variables nominales et numériques groupées par intervalles de classe. Ils sont composés d'un axe et d'une série de bandes horizontales ou verticales. Les bandes montrent les fréquences de différentes valeurs ou simplement les différentes valeurs elles-mêmes. Les nombres sur l'axe des x d'un diagramme à bandes horizontales ou sur l'axe des y d'un diagramme à bandes horizontales sont appelés l'échelle.

Lorsque vous créez un diagramme à bandes, 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.

Il existe trois types de diagrammes pour montrer les données d'une série chronologique :

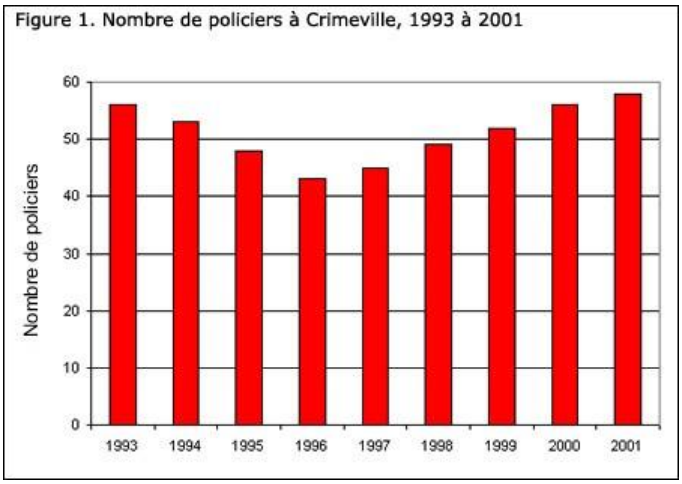
- diagrammes à bandes horizontales;
- diagrammes à bandes verticales;
- diagrammes linéaires.

Ces trois types de diagrammes fonctionnent bien pour comparer des valeurs. Cependant, en général, les comparaisons sont mieux présentées verticalement.

Exemple 1 – Diagrammes à bandes verticales

Les diagrammes à bandes doivent être utilisés lorsqu'on veut présenter des segments d'information. À l'aide de l'information dans la section sur les types de diagrammes, vous verrez que les diagrammes à bandes verticales sont 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. En un coup d'œil, vous pouvez voir que les échelles des axes augmentent en s'éloignant de l'origine. La figure 1 ci-dessous montre le nombre de policiers à Crimeville entre 1993 et 2001.

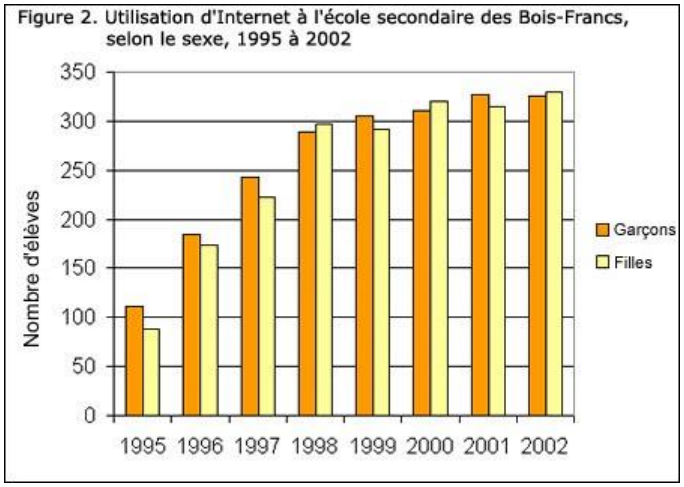
Figure 1. Nombre de policiers à Crimeville, 1993 à 2001



Dans la figure 1, vous pouvez voir que le nombre de policiers a diminué entre 1993 et 1996, mais qu'il a recommencé à croître en 1996. Le diagramme permet également de comparer facilement le nombre de policiers entre les années. Par exemple, en 2001, il y avait neuf policiers de plus qu'en 1998.

Le diagramme à bandes verticales doubles est un autre moyen efficace de comparer des ensembles de données sur les mêmes endroits ou éléments. Ce type de diagramme à bandes verticales donne au moins deux informations pour chacun des éléments sur l'axe des x au lieu d'une seule comme dans la figure 1. Vous pouvez faire des comparaisons directes dans un diagramme selon l'âge, le sexe, la race ou tout autre élément que vous désirez comparer. Cependant, si un diagramme à bandes verticales doubles comporte trop de séries de données, le diagramme devient encombré et il peut devenir difficile à lire.

La figure 2, un diagramme à bandes verticales doubles, compare deux séries de données : le nombre de garçons et de filles qui utilisent Internet à l'école secondaire des Bois-Francs entre 1995 et 2002. Une des barres représente le nombre de garçons et l'autre, le nombre de filles qui utilisent Internet.



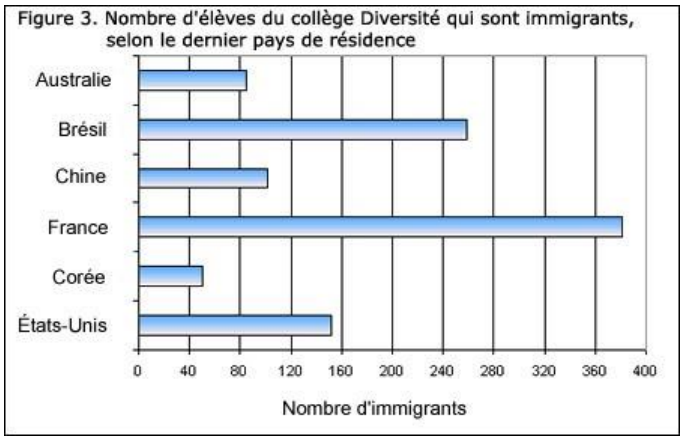
Un désavantage des diagrammes à bandes verticales, cependant, est le manque d'espace pour les étiquettes sous chacune des barres. Lorsque les étiquettes des catégories sont trop longues, il se peut qu'un diagramme à bandes horizontales soit préférable pour présenter l'information.

Exemple 2 – Diagrammes à bandes horizontales

Le diagramme à bandes horizontales utilise l'axe des y (ligne verticale) pour les étiquettes. Il y a plus d'espace pour les étiquettes des variables nominales sur l'axe des y.

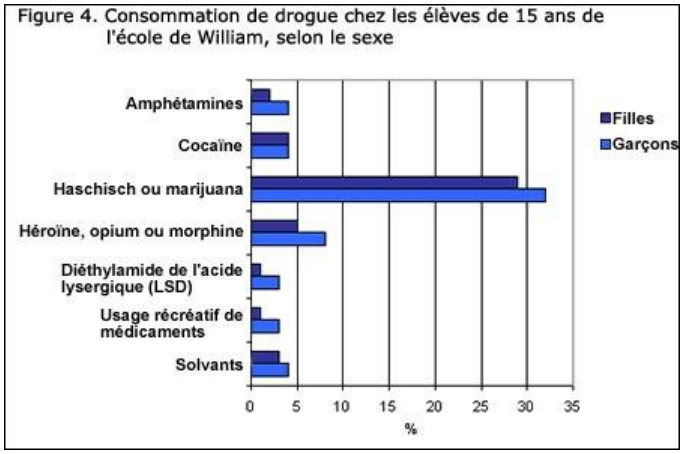
La figure 3 montre le nombre d'élèves du collège Diversité qui sont des immigrants selon le dernier pays de résidence. Vous pouvez constater que 100 élèves proviennent de la Chine, 380 de la France et 260 du Brésil.

Un diagramme à bandes horizontales a été utilisé pour montrer la comparaison des données. Ce diagramme est la meilleure façon de présenter ce type d'information parce que les étiquettes (dans le cas présent, le nom des pays) sont trop longues pour être lisibles sur l'axe des x.



Un diagramme à bandes horizontales doubles est semblable à un diagramme à bandes verticales doubles et il peut être utilisé lorsque les étiquettes sont trop longues pour être inscrites à l'axe des x.

Dans la figure 4, plus d'une information est transmise au public : la consommation de drogue chez les garçons de 15 ans est comparée à celle des filles du même âge à l'école de William. En incluant ces deux informations dans le même diagramme, on peut plus facilement faire une comparaison. Le diagramme montre que 32 % des garçons et 29 % des filles ont déjà consommé du haschisch ou de la marijuana et que 3 % des garçons et 1 % des filles ont déjà consommé du LSD. Le diagramme montre également qu'un pourcentage égal de garçons et de filles (4 %) ont déjà consommé de la cocaïne.



Exemple 3 – Comparer plusieurs endroits ou éléments

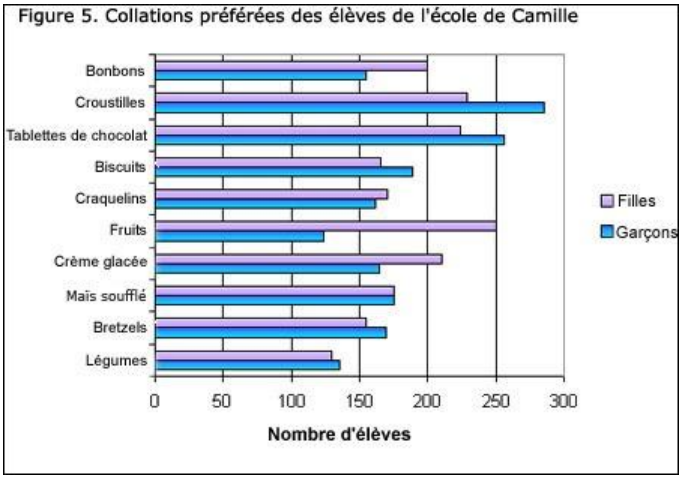
La figure 5 est un exemple d'un diagramme à bandes horizontales doubles. Camille a demandé à un nombre égal de garçons et de filles de son école secondaire de choisir leur collation préférée parmi les suivantes :

- maïs soufflé
- croustilles
- tablettes de chocolat
- craquelins
- bretzels
- biscuits
- crème glacée
- fruits
- bonbons
- légumes.

Elle a créé un diagramme pour montrer les résultats de son enquête. Regardez la figure 5 et répondez aux questions suivantes :

1. Quelle comparaison peut-on faire à l'aide de ce diagramme?
2. Quelle collation était la moins appréciée des filles?
3. Quelle collation était préférée de beaucoup plus de garçons que de filles?
4. Quelle collation était préférée de plus de filles que de garçons?
5. Quelle collation était appréciée d'autant de garçons que de filles?

Figure 5. Collations préférées des élèves de l'école de Camille



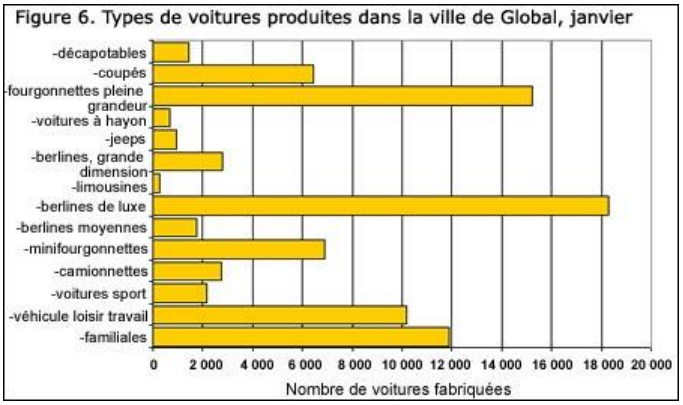
Réponses

1. Le diagramme montre une comparaison des collations préférées selon le sexe.
2. Les légumes constituent la collation la moins appréciée chez les filles.
3. Beaucoup plus de garçons préfèrent les croustilles.
4. Les filles aiment plus les bonbons, les craquelins, les fruits et la crème glacée que les garçons.
5. Un nombre égal de garçons et de filles préfèrent le maïs soufflé comme collation.

Exemple 4 – Utilisation inappropriée des diagrammes à bandes

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

Un diagramme à bandes horizontales pourra être plus efficace qu'un graphique linéaire lorsqu'il y a peu de périodes ou d'éléments. Si vous voulez comparer plus de 9 ou 10 éléments, il serait préférable d'utiliser un diagramme linéaire. La figure 6 est un exemple de situation où il serait préférable d'employer un graphique linéaire plutôt qu'un diagramme à bandes horizontales.



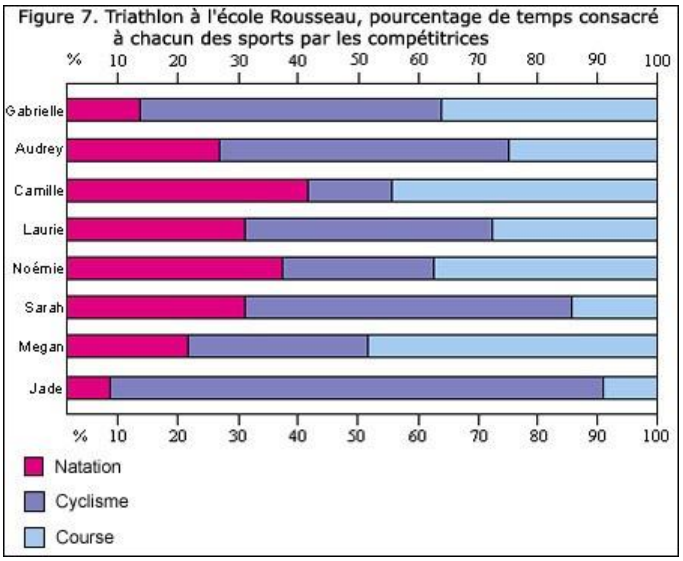
Exemple 5 – Autres diagrammes à bandes

Il y a plusieurs autres types de diagrammes à bandes qui peuvent être employés. La pyramide des âges est une application spéciale du diagramme à bandes doubles. Les exemples suivants sont rarement utilisés, mais ils peuvent être utiles si employés correctement.

Diagrammes à bandes superposées

Le diagramme à bandes superposées est un outil d'analyse préliminaire des données utilisé pour montrer les segments d'un tout. Statistique Canada les emploie rarement, même si ce type de diagramme peut contenir beaucoup d'information. Le diagramme à bandes superposées 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 la figure 7, 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 diagramme avait été utilisé pour représenter les données concernant un décathlon (10 épreuves), les données auraient été beaucoup plus difficiles à analyser.



Une autre raison pourquoi ces diagrammes ne sont pas employés souvent est qu'ils peuvent montrer un autre aspect de l'information que celui désiré. Dans l'exemple ci-dessus, Sarah a peut-être pris deux heures pour faire le triathlon et Laurie, trois, mais elles ont consacré le même pourcentage de leur temps à la première épreuve. Les deux ont consacré 30 % de leur temps à la natation, mais Laurie y a consacré 54 minutes et Sarah, 36 minutes. En d'autres termes, ce diagramme ne nous indique pas le classement, seulement le pourcentage de leur temps consacré à chacun des sports. Cette situation peut porter à confusion si on ne lit pas le graphique correctement.

Lignes directrices concernant les diagrammes à bandes horizontales, verticales et superposées

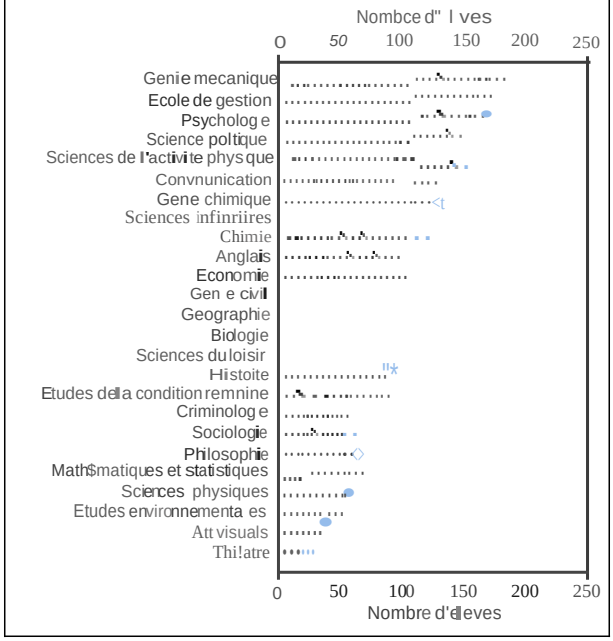
Vous devez vous souvenir des directives suivantes lorsque vous créez vos diagrammes à bandes :

- La largeur des bandes doit être plus grande que l'espace entre elles.
- Ne placez pas les lignes de référence en avant-plan.
- N'utilisez qu'un seul caractère pour le texte dans le diagramme. Essayez de garder la même police de caractère entre les diagrammes d'une même présentation ou d'un même document. Les polices linéales (sans empattement) sont préférables.
- L'ordre des couleurs devrait aller de la plus foncée à la plus pâle.
- Évitez les couleurs ou le remplissage trop voyants.

Diagrammes par points

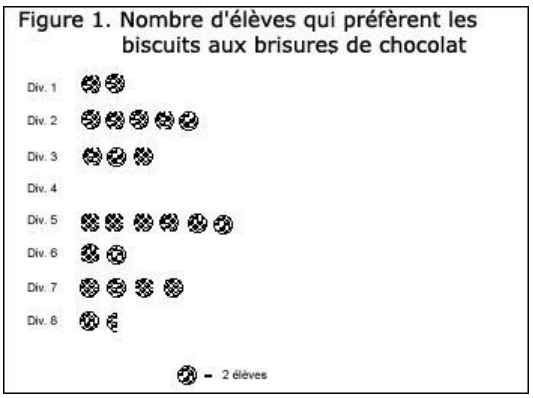
Un graphique par points est une des façons les plus simples de présenter de l'information de façon visuelle, mais il s'agit du diagramme le moins employé. La figure 8 est un exemple de graphique par points. Comme vous pouvez le voir, le message et l'information liés au diagramme peuvent être transmis rapidement et sont faciles à comprendre pour le lecteur.

Figure 8.Taux d'Inscription par programme,unlversite Academia



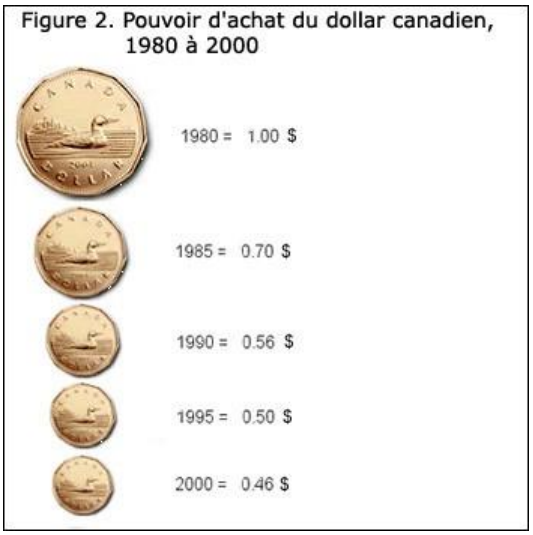
Pictogramme

Un pictogramme emploie des symboles pour donner la signification de l'information statistique. Ils devraient être employés avec soin parce qu'ils peuvent, accidentellement ou volontairement, donner une fausse représentation des données. C'est pourquoi un diagramme doit être précis en ce qui a trait à l'aspect visuel.



La figure 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, un demi biscuit représente donc un élève. Ces données auraient pu facilement être présentées dans un histogramme où l'échelle aurait servi à présenter les données plutôt qu'un symbole.

Voici un autre exemple de pictogramme.



La figure 2 montre comment le dollar canadien a diminué jusqu'à une valeur de 46,17 cents en 20 ans en raison de l'inflation. Cette information signifie que la valeur du dollar canadien de 2000 est moins de la moitié de celui de 1980! Quel est le problème avec la présentation des statistiques dans le pictogramme de la figure 2?

La taille des images (surface totale) des dollars (huards) dans le diagramme peut porter à confusion. La différence entre la valeur des dollars représentés est exagérée par les figures. Les diagrammes devraient refléter le pouvoir d'achat réel des dollars pour les années visées. Puisque 46 cents est tout juste inférieur à la moitié d'un dollar, le huard pour 2000 devrait sembler être tout juste plus petit que la moitié de celui de 1980; il devrait donc être deux fois plus gros que ce qu'il est.

On peut prétendre que les personnes qui ont regardé le diagramme 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 font 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. Statistique Canada utilise rarement les pictogrammes pour présenter l'information statistique, mais les médias les emploient très fréquemment.

Diagramme circulaire

Un diagrammecirculaire est une façon de résumer un ensemble de données nominales ou pour présenter les différentes valeurs d'une variable donnée (p.ex., répartition en pourcentage). Ce type de diagramme est formé d'un cercle divisé en segments. Chaque segment représente une catégorie particulière. La surface de chacun des segments représente la même proportion du cercle que la catégorie face à l'ensemble des données.

Les diagrammes circulaires montrent habituellement la partie d'un tout. Souvent, un segment sera séparé du reste du cercle afin de souligner l'importance de l'information.

Le diagramme circulaire montre que 90 % des élèves et des membres de la faculté de l'école Avenue ne désirent pas qu'un code vestimentaire soit instauré et que seulement 10 % des personnes à l'école désirent l'instauration d'un code vestimentaire. Ce fait est clairement souligné par la séparation du reste du cercle.

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

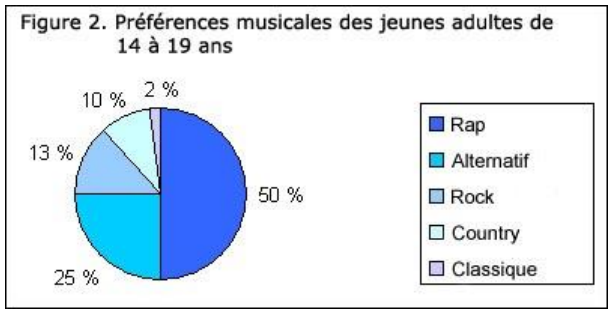
Le diagramme circulaire utilise les pourcentages pour comparer l'information. Les pourcentages sont utilisés parce qu'il est plus facile de représenter un tout de cette façon. Le tout est égal à 100 %. Par exemple, si vous passez sept heures à l'école et que 55 minutes de ce temps sont consacrées à votre dîner, alors 13,1 % de votre journée à l'école est consacrée au dîner. Pour présenter cette information dans un diagramme 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 diagramme 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 diagramme circulaire

Un diagramme circulaire est construit en convertissant la part de chacune des séries en un pourcentage de 360 degrés. Dans la figure 2, on peut bien voir les préférences musicales des 14 à 19 ans.



Le diagramme sectoriel 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 %).

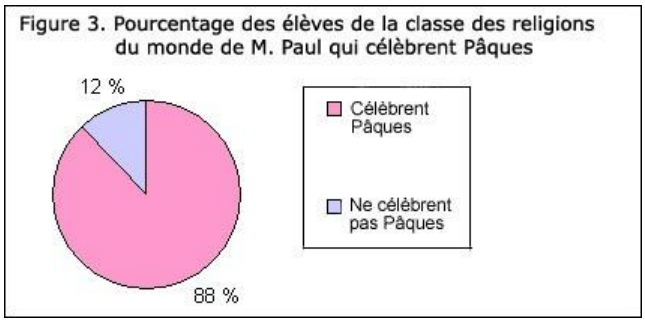
Astuce! *Lorsque vous créez un diagramme sectoriel, assurez-vous que les segments sont en ordre de grandeur (du plus grand au plus petit) dans le sens horaire.*

Afin de reproduire le diagramme sectoriel ci-dessus, suivez les étapes suivantes :

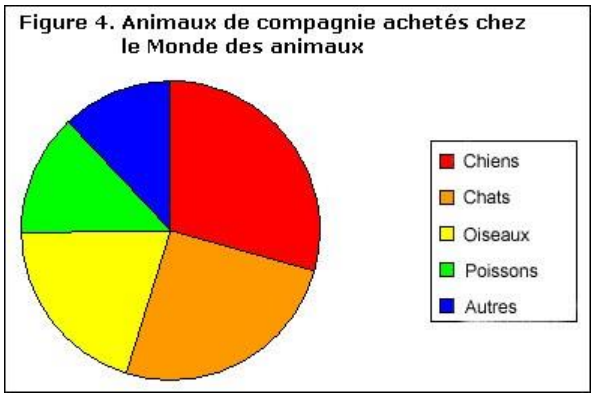
Si 50 % des élèves préfèrent le rap, 50 % du diagramme (360 degrés) équivaudraient à 180 degrés.

1. Dessinez un cercle avec votre rapporteur d'angles.
2. 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.
3. 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 diagramme. Si possible, le pourcentage et l'étiquette de la catégorie devraient être placés à côté du segment. De cette façon, les utilisateurs n'ont pas à revenir constamment à la légende pour savoir quelle catégorie est représentée par telle couleur.

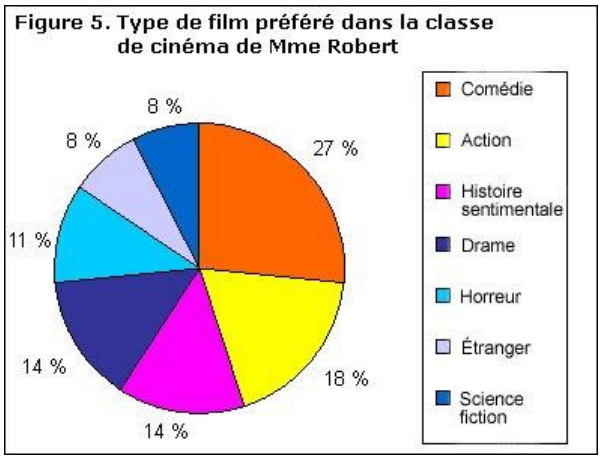


Ce diagramme circulaire transmet un message clair à l'utilisateur — 88 % des élèves de la classe de religions du monde célèbrent Pâques. On peut facilement comprendre le message simplement en regardant les pourcentages. Malheureusement, les étiquettes des catégories sont trop longues pour être inscrites à côté des segments : elles ont donc été placées en légende. Il est préférable de les placer à côté des segments.



Il est plus difficile de comprendre le message lié à la figure 4 parce qu'il n'y a pas de pourcentage lié aux segments. C'est pourquoi il est important d'inscrire les valeurs avec les segments.

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 segments du diagramme avaient été accompagnés d'une étiquette.



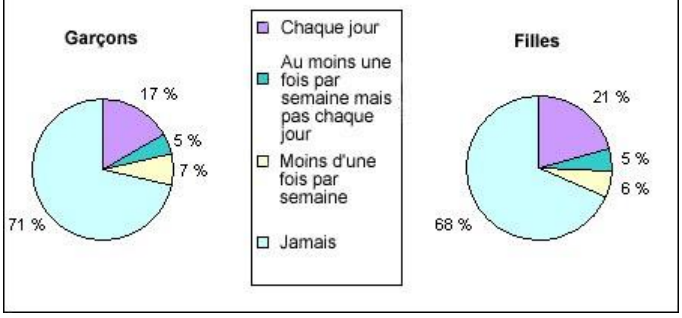
Dans le diagramme circulaire ci-dessus, la légende est bien construite et les pourcentages accompagnent les segments. Cependant, il y a trop d'éléments dans le diagramme pour qu'il soit possible d'avoir une idée rapide de la répartition des préférences. S'il y a plus de cinq ou six catégories, songez à utiliser un autre type de diagramme pour présenter l'information. Dans la figure 5, il aurait certainement été préférable d'utiliser un diagramme à bandes.

Astuce! *De nombreux logiciels peuvent dessiner les diagrammes circulaires rapidement et facilement. Cependant, les recherches ont montré que beaucoup de gens pouvaient se tromper en lisant un diagramme circulaire. En règle générale, les diagrammes à bandes peuvent communiquer le même message et ils risquent moins de créer une confusion.*

Diagrammes circulaires et diagrammes à bandes

Lorsque vous présentez de l'information statistique, n'utilisez pas plus d'un diagramme sectoriel.

Figure 6. Tabagisme chez les membres de l'équipe d'athlétisme de 15 ans de l'école secondaire du Parc



La figure 6 montre deux diagrammes circulaires côte à côte, alors qu'un diagramme à bandes double (deux barres côte à côte) aurait mieux présenté l'information. Un utilisateur pourrait éprouver des difficultés à comparer les segments d'un diagramme à ceux de l'autre. Par contre, dans un diagramme à bandes doubles, les segments deviennent des barres mises l'une à côté de l'autre, ce qui facilite la comparaison.

La figure 7 montre comment un diagramme à bandes doubles serait un meilleur choix pour présenter l'information comparativement à deux diagrammes circulaires. Le point principal pour la préparation de ce type de diagramme est de s'assurer d'utiliser la même échelle pour les deux côtés du diagramme. Vous verrez que l'information est beaucoup plus claire dans la figure 7 que dans la figure 6.

Figure 7. Revenus dans Utopie, selon le sexe

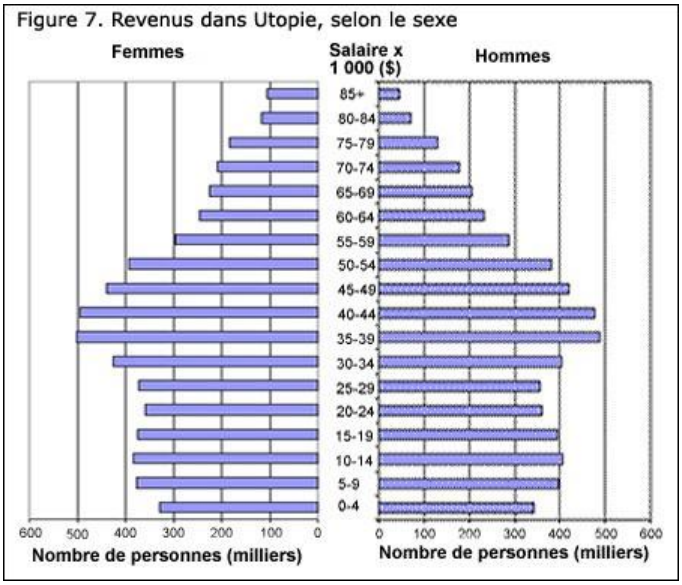


Diagramme linéaire

Les diagrammes linéaires, surtout employés en statistique et en science, sont utilisés plus fréquemment que tous les autres types de diagrammes ensembles, parce que leurs caractéristiques visuelles révèlent clairement les tendances dans les données de façon claire et qu'il s'agit d'un type de diagramme facile à créer.

Un diagramme 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. Pour de l'information sur la forme des diagrammes linéaires, voir le chapitre sur l'Organisation des données.

Les diagrammes 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 diagramme 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 diagramme linéaire est souvent considéré comme un diagramme de séries chronologiques. Par exemple, si vous vouliez montrer dans un diagramme 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.

Même s'ils ne présentent pas les données spécifiques aussi bien que les tableaux, les diagrammes linéaires peuvent montrer les relations plus clairement que les tableaux. Les diagrammes linéaires peuvent également montrer de nombreuses séries et conviennent donc habituellement mieux aux séries chronologiques et aux distributions statistiques.

Les diagrammes à bandes et à colonnes et les diagrammes linéaires ont un objectif commun. Le diagramme à colonne présente un changement d'amplitude, alors que le diagramme linéaire est employé pour montrer un changement d'orientation.

En résumé, les diagrammes linéaires :

- montrent bien les valeurs de données spécifiques révèlent
- les tendances et les relations entre les données comparent
- les tendances de différents groupes de variables

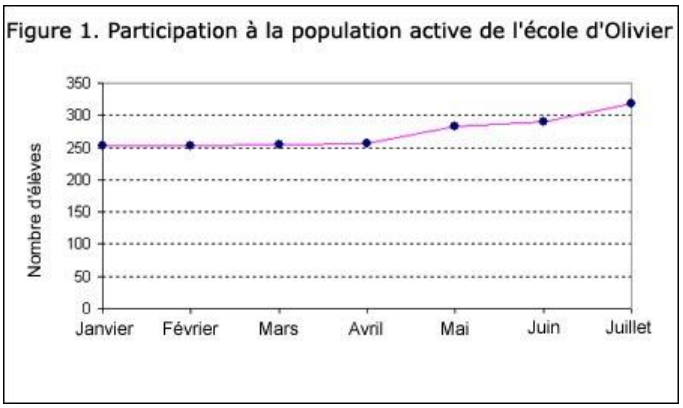
Les diagrammes peuvent donner une figure déformée des données. Si les échelles employées pour les axes d'un diagramme font paraître les données d'une certaine façon, alors le diagramme peut faire ressortir une tendance autre que celle désirée. Cela signifie que les intervalles entre les points adjacents d'un axe peuvent être inégaux ou que la même donnée, placée dans deux échelles différentes, semblera différente.

Exemple 1 – Extraire une tendance avec le temps

La figure 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 échelonné à l'aide d'intervalles sur l'axe des y, alors que la variable du temps se trouve à l'axe des x.

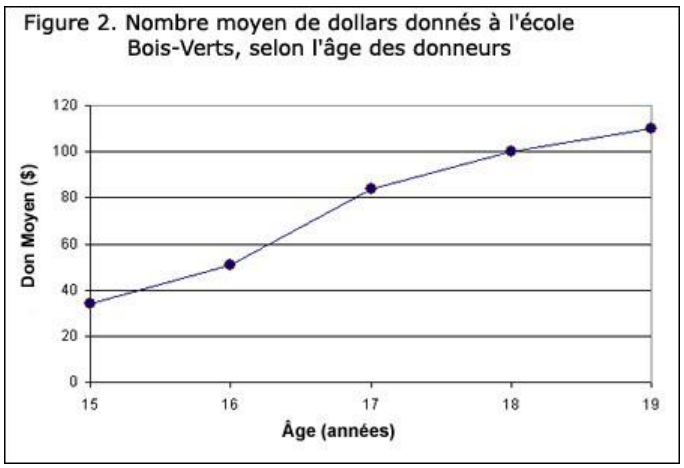
Le nombre d'élèves qui font partie de la population active était de 252 en janvier, 252 en février, 255 en mars, 256 en avril, 282 en mai, 290 en juin et 319 en juillet. En examinant le diagramme plus attentivement, on s'aperçoit que la participation des élèves était constante pour les quatre premiers mois du diagramme (janvier à avril) et que pour les trois autres mois (mai à juillet) le nombre s'est accru de façon constante.

Figure 1. Participation à la population active de l'école d'Olivier



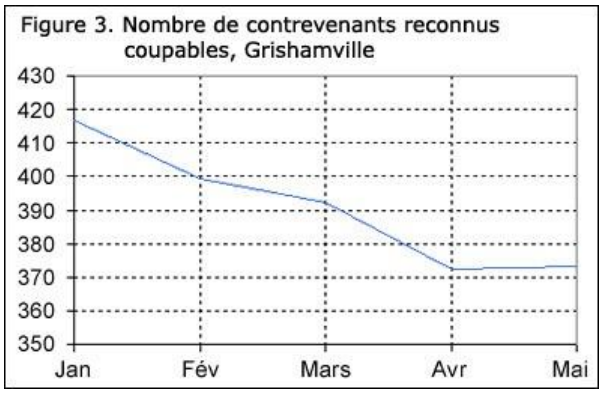
Exemple 2 – Comparer deux variables liées

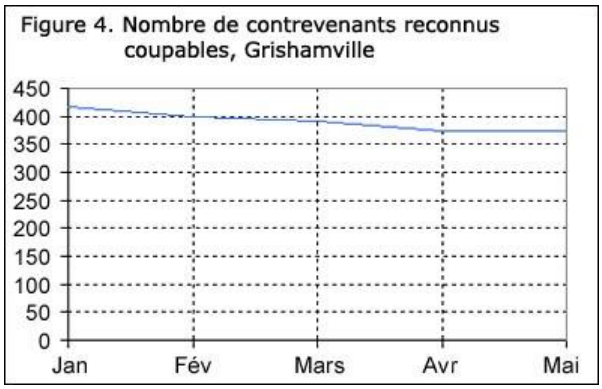
La figure 2 est un diagramme linéaire simple qui compare deux éléments; dans cette situation, le temps n'est pas un facteur. Le diagramme compare le nombre de dollars donnés selon l'âge des donneurs. Selon la tendance dans le diagramme, plus le donneur est âgé, plus il donne un montant important. Le donneur de 17 ans donne, en moyenne, 84 $. Chez les donneurs âgés de 19 ans, le don moyen est supérieur de 26 $, soit 110 $.



Exemple 3 – Utiliser la bonne échelle

Lorsque vous tracez une ligne, il est important d'utiliser la bonne échelle. Si vous ne le faites pas, la ligne pourra donner au lecteur une impression fausse quant aux données. Comparez la figure 3 et la figure 4 :



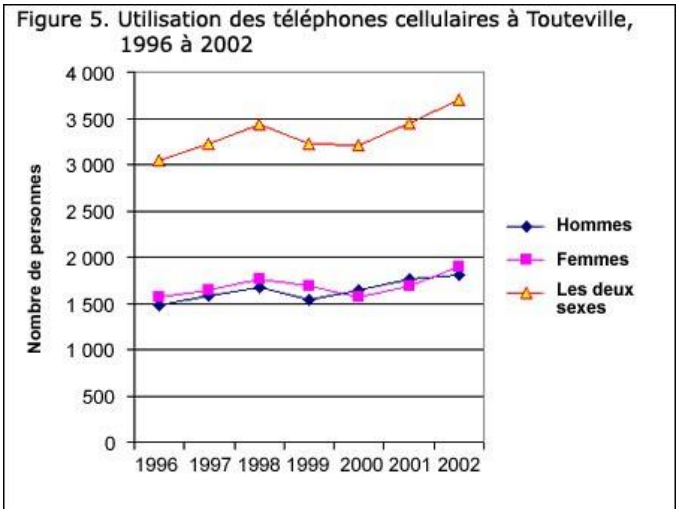


L'utilisation d'une échelle de 350 à 430 (figure 3) met l'accent sur un petit éventail de valeurs. Elle ne permet pas de montrer clairement la tendance en ce qui a trait aux contrevenants reconnus coupables entre janvier et mai, puisqu'elle exagère la tendance et qu'elle n'est pas liée à la figure générale. Cependant, le choix d'une échelle de 0 à 450 (figure 4) montre mieux à quel point le déclin des contrevenants reconnus coupables est réellement minime.

Exemple 4 – Diagrammes linéaires multiples

Les diagrammes linéaires multiples permettent de comparer efficacement des éléments semblables pour une même période (figure 5).

Figure 5. Utilisation des téléphones cellulaires à Touteville, 1996 à 2002



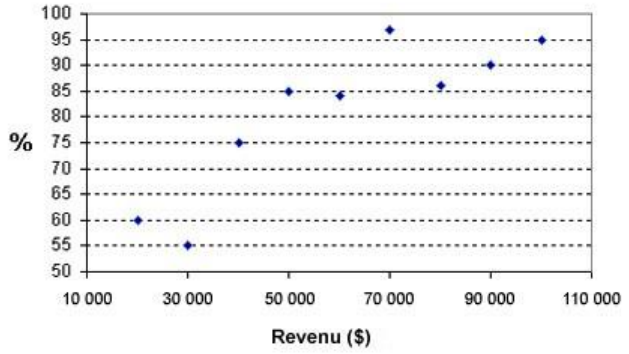
La figure 5 est un exemple d'un très bon diagramme. Le message est énoncé clairement dans le titre et chaque ligne du diagramme est identifiée correctement. Il est facile de voir dans ce diagramme que l'utilisation totale des téléphones cellulaires s'est accrue de façon constante depuis 1996, sauf pendant une période de deux ans (1999 et 2000) 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.

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 variables de l'axe des y dépendent des valeurs de la variable de l'axe des x (habituellement une variable indépendante).

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 (voir la figure 1 ci-dessous).

Figure 1. Possession d'une voiture, à Touteville, selon le revenu



Le pourcentage de gens qui possèdent une voiture augmente avec le revenu, ce qui montre une relation positive entre ces variables.

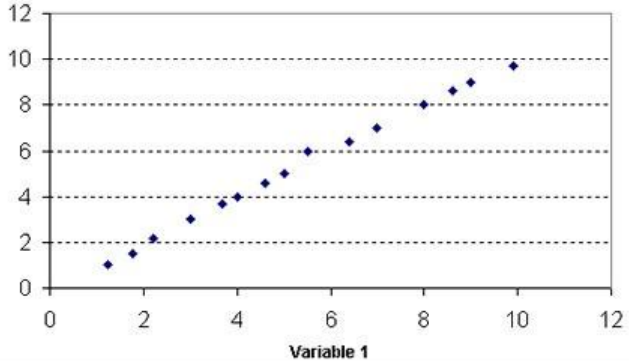
La tendance des points du nuage montre la relation entre les variables. Les nuages de points peuvent montrer différentes tendances et relations, par exemple :

- dépendance statistique
- relation positive ou directe entre les variables
- relation négative ou inverse entre les variables
- points dispersés
- tendances non linéaires
- répartition des données
- valeurs aberrantes.

Dépendance statistique

Lorsque les points forment une ligne droite dans le diagramme, la relation linéaire entre les variables est plus forte, tout comme la dépendance (figure 2).

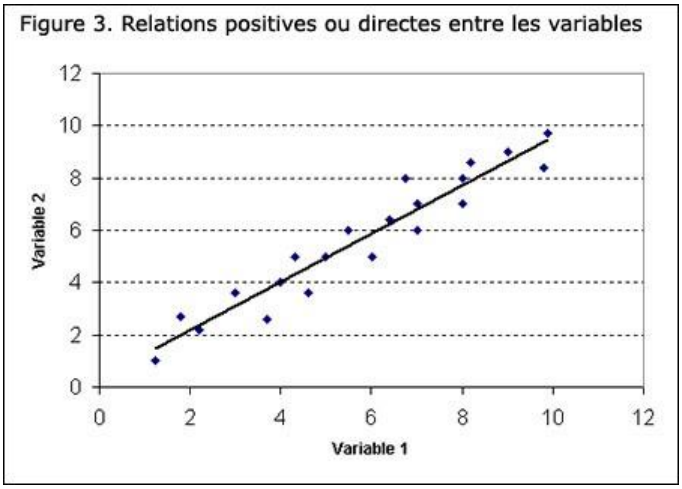
Figure 2. Relation linéaire forte entre les variables



Relations positives ou directes

Si les points sont regroupés près d'une ligne qui va du coin inférieur gauche au coin supérieur droit du diagramme, la relation entre les deux variables est positive ou directe (figure 3). Une augmentation de la valeur de x a de fortes chances d'être liée à une augmentation de la valeur de y. Plus les points sont en ligne, plus la relation est forte.

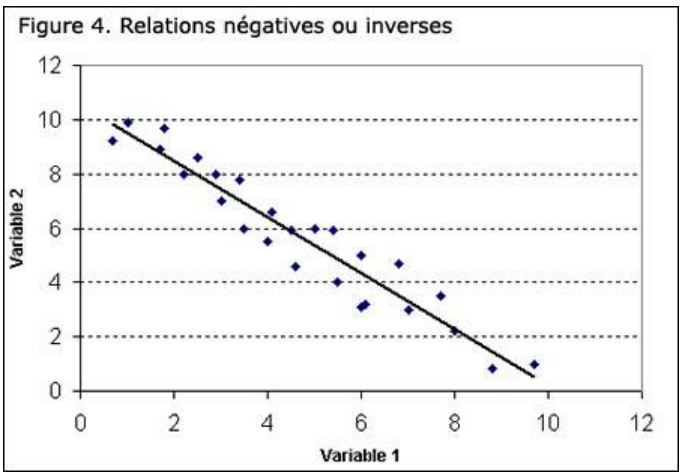
Figure 3. Relations positives ou directes entre les variables



Relations négatives ou inverses

Si les points sont regroupés près d'une ligne qui va du coin supérieur gauche au coin inférieur droit du diagramme, la relation entre les variables est négative ou inverse (figure 4).

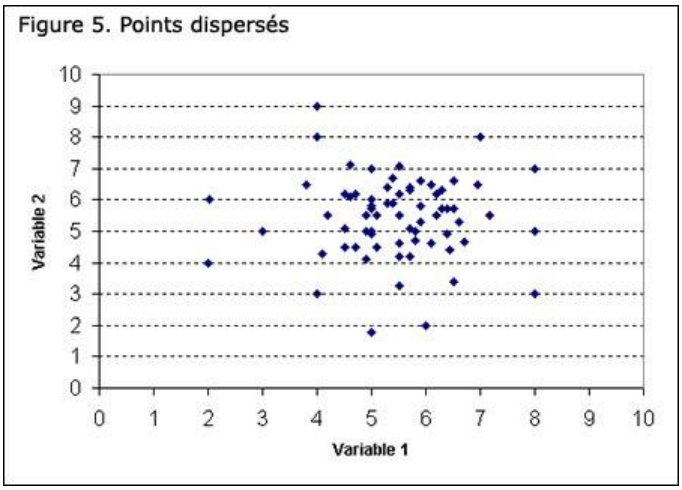
Figure 4. Relations négatives ou inverses



Points dispersés

Si les points sont dispersés de façon aléatoire, il n'y a pas de relations entre les variables; cela signifie que la dépendance entre les variables est faible ou nulle (figure 5).

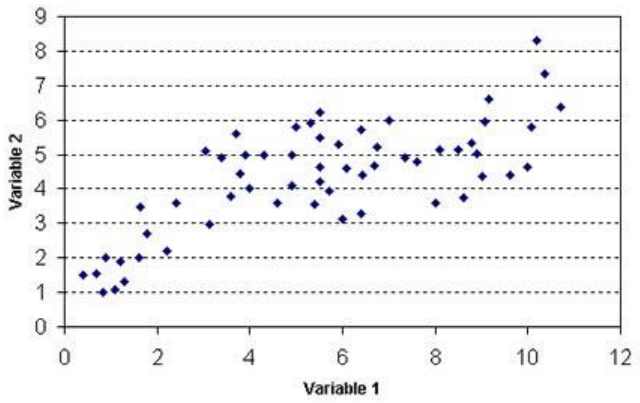
Figure 5. Points dispersés



Tendances non linéaires

Une dépendance très faible ou nulle peut être due à une relation non linéaire entre les variables. Si la relation est bien non linéaire (c.-à-d., les points sont regroupés en forme de courbe au lieu d'une ligne directe), le coefficient de corrélation linéaire ne sera pas une bonne mesure de la force de la relation (figure 6).

Figure 6. Dépendance très faible ou nulle



Répartition des données

Un nuage de points montrera également si les données sont grandement réparties ou si elles sont concentrées dans un secteur (figures 7 et 8).

Figure 7. Données grandement dispersées

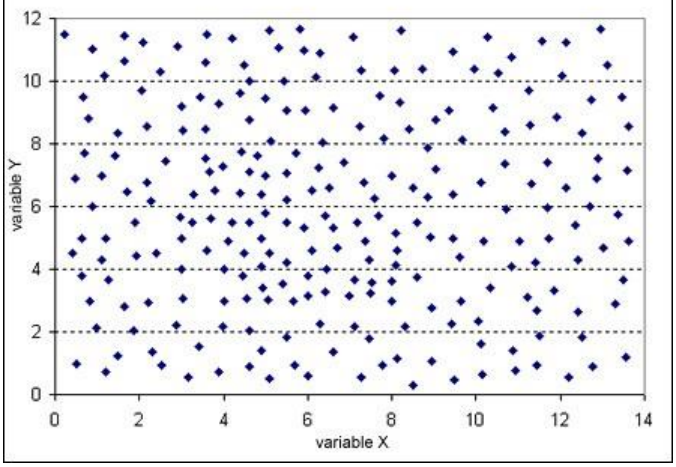
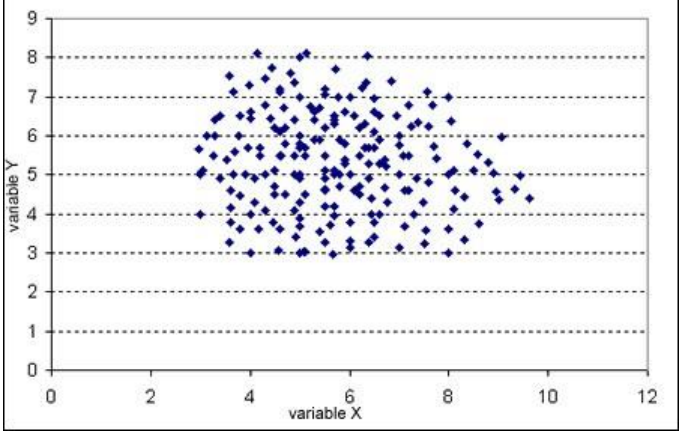


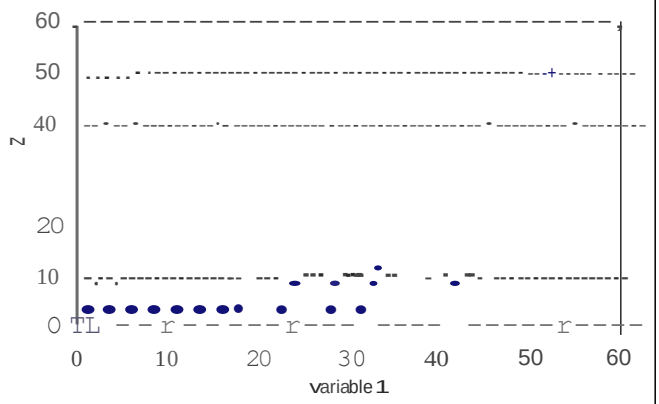
Figure 8. Données concentrées



Valeurs aberrantes

En plus de montrer une relation non linéaire entre deux variables, un nuage de points peut également montrer si des valeurs aberrantes existent (figure 9).

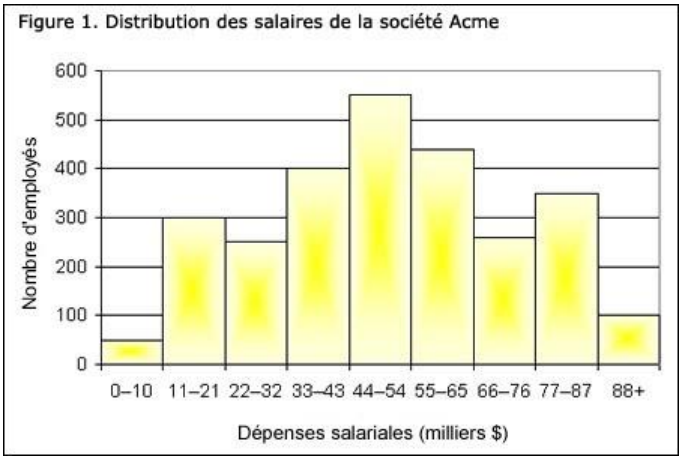
Figure 9. Valeurs aberrantes



Histogramme et polygone statistique

L'histogramme est un outil fréquemment utilisé, notamment pour résumer des données discrètes ou continues mesurées dans une échelle d'intervalle. Il est souvent employé pour montrer les caractéristiques principales de la distribution des données de façon pratique. Un histogramme sépare les valeurs possibles des données en classes ou groupes. Pour chaque groupe, un rectangle est construit dont la base correspond aux valeurs de ce groupe, et dont la taille du rectangle est proportionnelle au nombre d'observations dans le groupe. Cela signifie que les rectangles seront d'une hauteur différente. Un histogramme a une apparence semblable au diagramme à bandes verticales, mais lorsque les variables sont continues, il n'y a pas d'écart entre les barres. Lorsque les variables sont discrètes, des écarts devraient être laissés entre les barres. La figure 1 est un bon exemple d'histogramme.

Figure 1. Distribution des salaires de la société Acme



Un diagramme à bandes verticales diffère d'un histogramme de par les éléments suivants :

- Dans un histogramme, la fréquence est mesurée par la surface de la colonne.
- Dans un diagramme à bandes verticales, la fréquence est mesurée par la hauteur de la barre.

Caractéristiques de l'histogramme

En règle générale, l'histogramme possède des barres d'une largeur égale, même si ce n'est pas le cas lorsque les intervalles des classes sont de tailles différentes. Choisir la bonne largeur de barre pour l'histogramme est important. Comme vous pouvez le voir dans l'exemple ci-dessus, l'histogramme consiste simplement en un ensemble de barres verticales. Les valeurs de la variable étudiée sont mesurées selon l'échelle arithmétique de l'axe des x. Les barres sont de largeur égale et correspondent à des intervalles de classes égales, alors que la hauteur des barres correspond à la fréquence de la classe qu'elle représente.

L'histogramme est utilisé pour les variables dont les valeurs sont numériques et mesurées dans une échelle d'intervalle. Il est habituellement employé pour de grands ensembles données (plus de 100 observations). Un histogramme peut également faciliter la détection d'observations inhabituelles (valeurs aberrantes) ou tout manque dans les données.

Polygones statistiques

Un polygone statistique est un diagramme formé en reliant les points médians des colonnes d'un histogramme. Ces diagrammes sont utilisés seulement pour présenter des données concernant des variables continues d'un histogramme.

Le polygone statistique aplanit les changements abruptes pouvant être visibles dans un histogramme. Il est utile pour montrer la continuité de la variable à l'étude. Les figures 2 et 3 sont de bons exemples de polygones statistiques.

Figure 2. Distribution des salaires de la société Acme

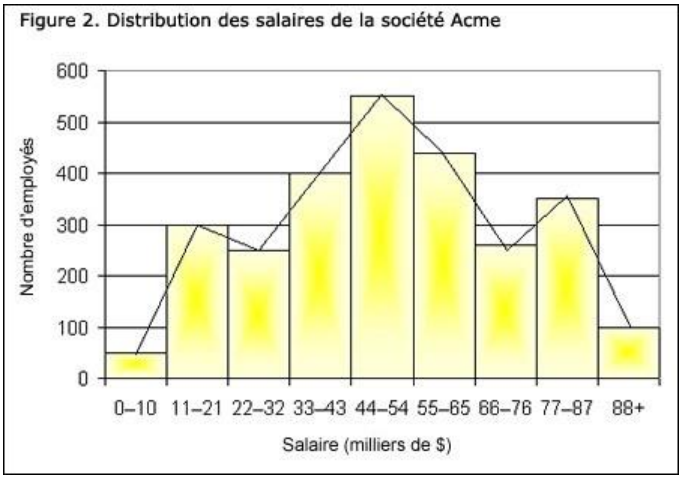
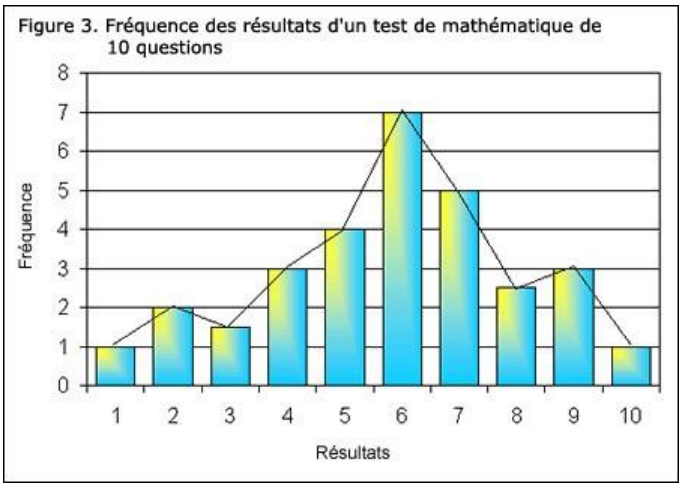


Figure 3. Fréquence des résultats d'un test de mathématique de 10 questions



Contrairement à la figure 2, ce polygone statistique comprend des espaces entre les barres. En regardant la figure, le lecteur peut immédiatement dire que les espaces signifient que les variables sont discrètes. De cette façon, les polygones statistiques permettent au lecteur de déterminer plus facilement le type de variables utilisées.

Résumé

Si vous décidez que le diagramme est le meilleur moyen de présenter votre information, peu importe le type de diagramme choisi, vous devez vous souvenir des 10 conseils suivants :

Dix conseils pour que vos diagrammes soient impressionnants!

1. choisissez un message important
2. ayez un objectif précis
3. mettez l'accent sur le message, pas sur la source
4. essayez différentes possibilités et différents types de diagrammes
5. utilisez un design simple pour les données complexes
6. faites « parler » les données
7. adaptez la présentation du diagramme au public cible
8. assurez-vous que la perception visuelle est facile et précise
9. évitez les déformations et les ambiguïtés
10. optimisez le design et intégrez le style dans le texte et les tableaux

Résumé

Types de diagrammes	Description
pyramide des âges	Représente la structure de l'âge d'une population.
diagramme à bandes verticales	Compare des valeurs importantes. Montre les données d'une meilleure façon que le diagramme à bandes horizontales et il est préférable de l'utiliser si possible.
diagramme par points	Montre un grand nombre de catégories lorsque l'ordre de ces dernières n'est pas important. Plus utile pour la présentation de valeurs de catégories en ordre décroissant.
histogramme	Montre les données de variables discrètes ou continues d'une façon similaire aux diagrammes à colonnes, mais sans l'écart entre les colonnes.
polygone statistique	Montre les données de variables continues. Aplanit les changements abrupts pouvant être visibles dans un histogramme.
diagramme à barres horizontales	Compare des données importantes. Utile lorsque le nom des catégories est trop long pour être inscrit au bas de la colonne.
diagramme linéaire	Souvent utilisé pour montrer l'évolution des données sur une période précise.
pictogramme	Préféré des graphistes, mais les élèves peuvent également créer des présentations figuratives simples. Les comparaisons doivent être représentées de façon exacte et respecter l'échelle.
diagramme circulaire	Compare un nombre plus petit de catégories. Les valeurs doivent être inscrites comme étant différentes, ou les différences pourraient être difficiles à trouver. Apposer la valeur des segments dans le diagramme règle ce problème. Lorsque des points de données sont semblables, le message du diagramme circulaire pourra être mal compris. Un diagramme à bandes pourrait être plus approprié.
nuage de points	Mesure deux ou plusieurs variables qui semblent liées.

Créez votre propre diagramme

Maintenant vous êtes prêt à pratiquer la production de diagrammes simples. Utilisez l'outil au bas de la page pour vous aider à créer vos propres diagrammes. Vous n'avez qu'à suivre les directives ci-dessous :

1. Inscrivez le titre de votre diagramme dans l'espace sous l'étiquette *Titre du diagramme*. Vous pouvez y entrer jusqu'à 80 caractères. Par exemple, si vous voulez créer un diagramme à bandes pour montrer les données d'une enquête concernant les cours préférés de vos camarades. Vous pourriez intituler votre diagramme « Matières préférées, par nombre d'élèves ».

   Pour un diagramme ayant des étiquettes aux axes des x et des y, remplissez les espaces sous l'*Étiquette de l'axe des x* et l'*Étiquette de l'axe des y*. Vous pouvez inscrire jusqu'à 30 caractères dans chaque étiquette de catégorie.

2. À l'aide du même exemple, vous pourriez intituler l'axe des x « Matières » et l'axe des y « Nombre d'élèves ».

   Nota : Veuillez laisser le titre et les étiquettes de l'axe des x en blanc pour les diagrammes de type boîtes et tracés filiformes.

3. Dans l'espace *Étiquettes* des catégories de l'axe des x, inscrivez les différentes catégories que vous voulez voir dans l'axe des x ou les catégories des segments dans le cas d'un diagramme à segments. Inscrivez chacune des catégories l'une sous l'autre et n'utilisez pas de ponctuation entre les catégories.

   Dans l'exemple, les catégories pourraient être « Histoire, algèbre, biologie, anglais et théâtre ».

4. N'inscrivez rien dans l'espace *Premier ensemble de données – Légende*, à moins que vous désiriez comparer plusieurs types de données dans un diagramme. Si vous remplissez ces espaces, une légende accompagnera votre diagramme, indiquant au lecteur ce que chaque couleur représente.

   Par exemple, si vous avez décidé de comparer les matières préférées des garçons à celles des filles, vous devriez inscrire « Garçons » dans le *Premier champ des données – Légende* et « Filles » dans le *Deuxième champ des données – Légende*. Vous pourriez également inscrire un troisième ensemble regroupant les données des deux sexes, « Garçons et filles », sur les matières préférées.

5. Sous chaque boîte de la légende, une autre boîte sert à ajouter les pourcentages ou nombres liés aux catégories. Inscrivez les nombres correspondant aux étiquettes des catégories dans la boîte *Étiquettes des catégories de l'axe des x*.

   Dans l'exemple, sous « Histoire », inscrivez 35 pour les garçons et 27 pour les filles. Cela signifie que 35 garçons et 27 filles ont affirmé que l'histoire était leur matière préférée. Continuez ainsi jusqu'à ce que tous les nombres aient été entrés.

   Nota : Si vous avez choisi un diagramme circulaire, les nombres seront convertis en pourcentages. Si les nombres ne comptent pas pour 100 %, ils doivent être convertis en pourcentages de façon à ce que les segments totalisent 100 %.

6. Choisissez la couleur pour votre ensemble de données à partir du menu déroulant. Si vous faites un diagramme à bandes, vous devez choisir la couleur des barres. Si vous faites un diagramme à bandes doubles ou un diagramme linéaire, assurez-vous d'utiliser des couleurs différentes pour chaque ensemble de données. Si vous ne variez pas les couleurs, par défaut le premier ensemble de données sera rouge, le deuxième sera bleu et le troisième sera vert.

   Nota : Les couleurs d'un diagramme à segments sont choisies automatiquement par le logiciel et les boîtes et tracés filiformes utilisent une même couleur.

7. Cliquez sur les boîtes qui s'appliquent à votre diagramme. La boîte sous l'étiquette *Utilisez l'origine* (0,0) est réglée automatiquement. Cependant, si vous ne voulez pas que votre diagramme commence à l'origine, mais à partir de votre premier ensemble de données, cliquez sur la boîte pour retirer le crochet. Si vous faites un diagramme à segments, cliquez dans la boîte *Montrer* les étiquettes dans le diagramme circulaire du diagramme à segments. Si vous voulez montrer chaque point de votre diagramme, choisissez la boîte *Montrer* les points dans le diagramme.

8. Choisissez le type de diagramme que vous voulez créer dans le menu déroulant.

   Nota : Les *diagramme à deux dimensions* ressemble à un diagramme linéaire. Les points sont placés à l'aide d'une ligne. Cependant, tout ce qui se trouve sous la ligne est coloré comme les barres d'un diagramme à bandes. Ce type de diagramme montre quelle superficie est couverte en colorant l'espace. Ce type de diagramme peut être utilisé pour souligner une tendance (diminution importante, constance, etc.) dans les données.

9. Appuyez sur le bouton *Créer le diagramme* et votre diagramme apparaîtra à l'écran. Pour voir les données selon différents types de diagrammes, cliquez sur le bouton « Retour » et choisissez un autre type de diagramme dans le menu déroulant.

Souvenez-vous que lorsque vous utilisez cet outil, les valeurs de l'axe des y sont considérées comme numériques et que toute inscription non numérique est égale à 0. Les valeurs de l'axe des x sont considérées comme étant des étiquettes de catégories et se trouveront dans le diagramme comme tel. Sauf pour les

diagrammes de type boîtes et tracés filiformes, assurez-vous d'entrer l'étiquette de l'axe des x (c.-à-d., l'année, le mois, le jour, la profession, etc.) Si vous omettez d'entrer les valeurs textuelles dans l'axe des x, le résultat ne sera pas ce à quoi vous vous attendiez.

Titre du diagramme :

Étiquette de l'axe des x :

Étiquette de l'axe des y :

Premier ensemble de données – Légende :

Deuxième ensemble de données – Légende :

Troisième ensemble de données – Légende :

Étiquettes des catégories de l'axe des x :

Premier ensemble de données Points des données :

Deuxième ensemble de données Points des données :

Troisième ensemble de données Points des données :

Choisissez une couleur :

Premier ensemble de données Gris pâle ▼

Deuxième ensemble de données Bleu ▼

Troisième ensemble de données Jaune ▼

☑ Utiliser l'origine (0,0)
☐ Montrer les points dans le diagramme
☐ Montrer les étiquettes dans le diagramme circulaire

Type de diagramme : Diagramme linéaire ▼

Créer le diagramme

Exercice

1. Pendant une semaine, recueillez des exemples de diagrammes (c.-à-d., diagrammes à bandes, diagrammes circulaires, diagrammes linéaires) dans les journaux, les magazines et les revues scientifiques. À la fin de la semaine, regardez les exemples de diagrammes et déterminez s'ils sont appropriés ou si les statistiques auraient été mieux présentées dans un autre type de diagramme.

2. Quel type de diagramme utiliseriez-vous pour présenter les données suivantes? Expliquez vos choix.
    a. le nombre de filles dans chacune des années à votre école
    b. le nombre annuel de décès de la route dans votre province ou territoire au cours des 30 dernières années
    c. la vitesse (km/h) des 20 animaux les plus rapides
    d. la population du Canada, des provinces et territoires

3. Discutez en classe des différentes situations où un diagramme à bandes horizontales ou à barres verticales serait le plus utile.

Diagrammes analytiques

On peut présenter les variables numériques de diverses façons, par exemple sous forme de *diagrammes à tiges et à feuilles* ou de *tableaux de distribution de fréquences*, tel que présenté dans la section intitulée « Organisation des données ».

Dans la présente section, on traite de deux autres façons de présenter les variables numériques : les *diagrammes de fréquences cumulées* et les *diagrammes de pourcentages cumulés*. Ces diagrammes analytiques sont très utiles pour trouver la tendance centrale de grands ensembles de données, ainsi que d'autres renseignements importants.

Fréquence cumulée

On utilise la *fréquence cumulée* pour déterminer le nombre d'observations qui se situent au-dessus (ou au-dessous) d'une valeur particulière dans un ensemble de données. On calcule la fréquence cumulée à l'aide d'un tableau de distribution de fréquences, qu'on peut construire à partir de diagrammes à tiges et à feuilles ou directement à partir des données.

On calcule la fréquence cumulée en ajoutant chaque fréquence tirée d'un tableau de distribution de fréquences à la somme de celles qui précèdent. La dernière valeur sera toujours égale au total des observations, puisque toutes les fréquences auront déjà été ajoutées au total précédent.

Variables discrètes ou continues

On peut caractériser des variables à l'intérieur d'un calcul à l'aide de la valeur qui leur est attribuée. Une *variable discrète* se compose de catégories séparées et indivisibles. Il ne peut exister de valeur entre une variable et sa voisine. Par exemple, si vous deviez observer les présences en classe enregistrées d'un jour à l'autre, vous pourriez découvrir que la classe compte 29 élèves une journée et 30 une autre journée. Il est cependant impossible que le nombre de présences en classe se situe entre 29 et 30. (Il n'y a simplement pas de place pour observer une valeur entre ces deux variables, étant donné qu'il est absolument impossible d'avoir une classe comptant 29 élèves et demi.)

On ne caractérise pas toutes les variables comme étant des variables discrètes. Certaines (comme le temps, la taille et le poids) ne se limitent pas à un ensemble fixe de catégories indivisibles. On appelle ces variables des variables continues; ces dernières sont divisibles en un nombre infini de valeurs possibles. Par exemple, on peut mesurer le temps en parties fractionnaires d'heures, de minutes, de secondes et de millisecondes. Alors, plutôt que d'achever une course en 11 ou 12 minutes, un jockey et son cheval peuvent franchir la ligne d'arrivée en 11 minutes et 43 secondes.

Il est essentiel de connaître la différence entre les deux types de variables pour calculer correctement leur fréquence cumulée.

Exemple 1 – Variables discrètes

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.

1. Utilisez ces variables discrètes pour :
   - créer un diagramme à tiges et à feuilles et d'autres colonnes intitulées *Fréquence, Valeur supérieure et Fréquence cumulée;*
   - déterminer la fréquence des observations pour chaque tige;
   - trouver la valeur supérieure de chaque tige;
   - calculer la fréquence cumulée en additionnant les nombres dans la colonne Fréquence;
   - enregistrer tous les résultats du tracé.
2. Tracez un diagramme à l'aide de l'axedesy (ou de la ligne verticale) pour la fréquence cumulée et de l'axedesx (ou de la ligne horizontale) pour le nombre de grimpeurs.

Réponses :

1. Le nombre de grimpeurs varie de 4 à 65. Pour produire un diagramme à tiges et à feuilles, il vaut mieux grouper les données en intervalles de classe de 10.

   On peut situer chaque intervalle dans la colonne *Tige*. Les nombres dans cette colonne représentent le premier nombre de l'intervalle de classe. (La *tige* 0, par exemple, représente l'intervalle 0 à 9, la tige 1, l'intervalle de 10 à 19 et ainsi de suite.)

   La colonne *Feuille* sert à indiquer le nombre d'observations qui se situent à l'intérieur de chaque intervalle de classe. Dans la *tige* 2 (l'intervalle 20 à 29), par exemple, les trois observations, 23, 24 et 26, sont représentées sous la forme de 3, 4 et 6.

   La colonne *Fréquence* sert à indiquer le nombre d'observations qui se situent à l'intérieur d'un intervalle de classe. Dans la tige 5, par exemple, il y a 9 feuilles (ou observations), alors que dans la tige 1, il n'y en a que deux.

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

   - Premièrement, ajoutez le nombre tiré de la colonne *Fréquence* au nombre précédent. Dans la tige 0, par exemple, nous n'avons qu'une seule observation et aucun nombre qui précède. La fréquence cumulée est donc un.

     1 + 0 = 1

   - Dans la *tige* 1, cependant, il y a deux observations. Ajoutez ces deux observations à la fréquence cumulée précédente (un) et vous obtiendrez trois.

1 + 2 = 3

- Dans la *tige* 2, il y a trois observations. Ajoutez ces trois observations à la fréquence cumulée précédente (trois) et vous obtiendrez comme fréquence cumulée pour la tige 2 le total (six).

  3 + 3 = 6

- Continuez à effectuer ces calculs jusqu'à ce que vous ayez additionné tous les nombres dans la colonne *Fréquence*.

- Enregistrez les résultats dans la colonne *Fréquence cumulée.*

La colonne *Valeur supérieure* sert à indiquer l'observation (la variable) ayant la valeur la plus élevée dans chacun des intervalles de classe. Dans la tige 1, par exemple, les deux observations 8 et 9 représentent les variables 18 et 19. La valeur supérieure de ces deux variables est 19.

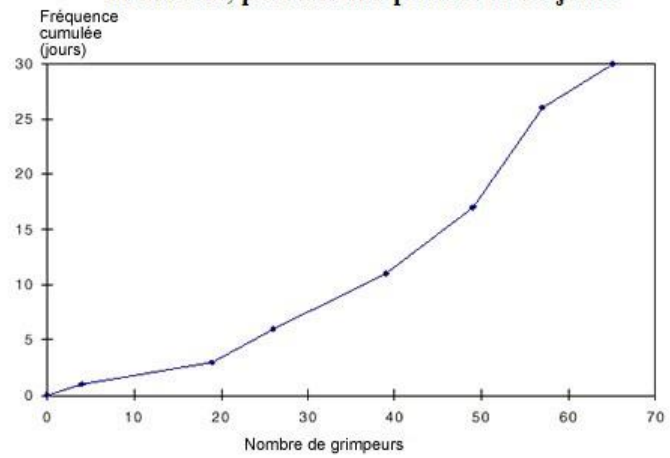
Tableau 1. Nombre de grimpeurs par jour autour du lac Louise, en Alberta, pendant une période de 30 jours

Tige	Feuille	Fréquence (f)	Valeur supérieure	Fréquence cumulée
0	4	1	4	1
1	8 9	2	19	1 + 2 = 3
2	3 4 6	3	26	3 + 3 = 6
3	1 5 5 7 9	5	39	6 + 5 = 11
4	0 1 2 3 5 9	6	49	11 + 6 = 17
5	0 1 1 2 4 4 5 6 7	9	57	17 + 9 = 26
6	0 2 3 5	4	65	26 + 4 = 30

2. Puisque ces variables sont des variables discrètes, utilisez les valeurs supérieures en traçant le diagramme. Tracez les points pour former une courbe continue appelée une ogive.

   Identifiez toujours le diagramme à l'aide de la fréquence cumulée, correspondant au nombre d'observations effectuées, sur l'axe vertical. Identifiez l'axe horizontal à l'aide de l'autre variable (le nombre de grimpeurs, dans ce cas) comme il est indiqué ci-dessous :

Figure 1. Nombre de grimpeurs par jour autour du lac Louise, en Alberta, pendant une période de 30 jours



Voici l'information qu'on peut obtenir à partir d'un diagramme ou d'un tableau :

- durant 11 des 30 jours, 39 personnes ou moins ont grimpé aux rochers entourant le lac Louise;
- durant 13 des 30 jours, 50 personnes ou plus ont grimpé aux rochers entourant le lac Louise.

Quand on utilise une variable continue, il faut adopter une méthode légèrement différente de celle utilisée pour une variable discrète afin de calculer la fréquence cumulée et de tracer le diagramme.

Exemple 2 – Variables continues

Voici l'épaisseur de la neige mesurée (arrondie au centimètre) et enregistrée pendant 25 jours à Whistler Mountain, en Colombie-Britannique :

242, 228, 217, 209, 253, 239, 266, 242, 251, 240, 223, 219, 246, 260, 258, 225, 234, 230, 249, 245, 254, 243, 235, 231, 257.

1. Utilisez les variables continues qui figurent ci-dessus pour :
   - créer un tableau de distribution de fréquences;
   - trouver la fréquence de chaque intervalle de classe;
   - situer l'extrémité de chaque intervalle de classe;
   - calculer la fréquence cumulée en additionnant les nombres dans la colonne Fréquence;
   - enregistrer tous les résultats à l'intérieur du tableau.

2. Utilisez l'information recueillie à partir du tableau de distribution de fréquences pour tracer un diagramme de fréquences cumulées.

Réponses :

1. Les mesures de l'épaisseur de la neige varient de 209 cm à 266 cm. Pour produire le tableau de distribution de fréquences, il vaut mieux grouper les données en intervalles de classe de 10 cm chacun.

   Chaque intervalle de classe de 10 cm compris entre 200 cm et 270 cm est indiqué dans la colonne *Épaisseur de la neige*.

   La colonne *Fréquence* sert à enregistrer le nombre d'observations qui s'inscrivent à l'intérieur d'un intervalle particulier. Cette colonne représente les observations inscrites dans la colonne Comptage, mais sous forme numérique seulement.

   La colonne *Extrémité* est en bonne partie la même que la colonne Valeur supérieure figurant dans l'exercice 1, sauf que l'extrémité est le nombre le plus élevé inclus à l'intérieur de l'intervalle, indépendamment de la valeur réelle de chaque observation. Dans l'intervalle de classe de 210 à 220, par exemple, la valeur réelle des deux observations est 217 et 219 respectivement. Toutefois, plutôt que d'utiliser 219, on utilise l'extrémité 220.

   La colonne *Fréquence cumulée* sert à indiquer le total de l'ajout de chaque fréquence à celle qui précède.

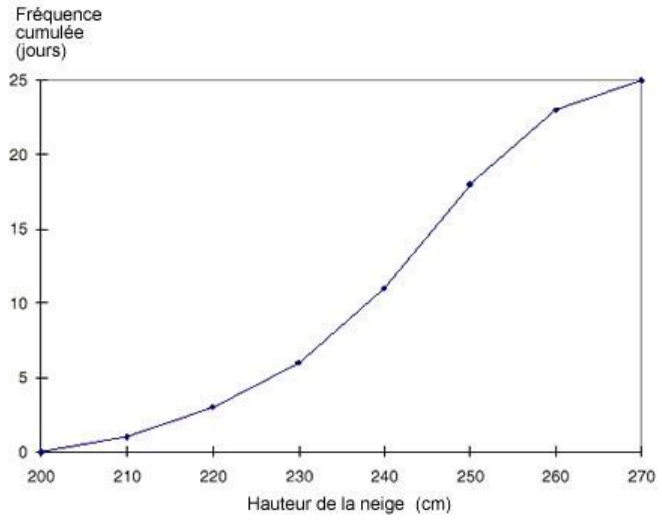
Tableau 2. Épaisseur de la neige mesurée à Whistler Mountain, en Colombie-Britannique, pendant une période de 25 jours

Épaisseur de la neige(x)	Comptage	Fréquence (f)	Extrémité	Fréquence cumulée
			200	0
200 à < 210	I	1	210	1
210 à < 220	II	2	220	3
220 à < 230	III	3	230	6
230 à < 240	𝍸	5	240	11
240 à < 250	𝍸 II	7	250	18
250 à < 260	𝍸	5	260	23
260 à < 270	II	2	270	25

2. Parce que la variable est continue, on utilise les extrémités de chaque intervalle de classe pour tracer le diagramme. On joint les points tracés pour former une ogive.

   Souvenez-vous que la fréquence cumulée (le nombre d'observations effectuées) est indiquée sur l'axe vertical y et que l'autre variable (l'épaisseur de la neige) est indiquée sur l'axe horizontal x, comme le montre la figure 2.

Figure 2. Épaisseur de la neige mesurée à Whistler Mountain, en Colombie-Britannique, pendant une période de 25 jours



Voici l'information qu'on peut obtenir à partir du diagramme ou du tableau :

- l'épaisseur de la neige n'a jamais été inférieure à 200 cm pendant une période de 25 jours;
- l'épaisseur de la neige a été inférieure à 210 cm une journée sur 25;
- l'épaisseur de la neige a été égale ou supérieure à 260 cm pendant deux des 25 jours.

Autres calculs de la fréquence cumulée

La *distribution de fréquences relatives* est un autre calcul qu'on peut effectuer à l'aide d'un tableau de distribution de fréquences. On définit cette méthode comme étant le pourcentage d'observations s'inscrivant à l'intérieur de chaque intervalle de classe. On peut trouver la fréquence relative cumulée en divisant la fréquence de chaque intervalle par le nombre total d'observations. (Pour plus de renseignements, voir Distributiondefréquences dans le chapitre intitulé *Organisation des données*.)

On peut aussi utiliser un tableau de distribution de fréquences pour calculer un *pourcentage cumulé*. Cette méthode de distribution de fréquences nous donne le pourcentage de la fréquence cumulée, au lieu de nous donner uniquement le pourcentage de la fréquence.

Pourcentage cumulé

Le pourcentage cumulé est une autre façon de représenter une distribution de fréquences. Il permet de calculer le pourcentage de la fréquence cumulée dans chaque intervalle, tout comme une distribution de fréquences relatives permet de calculer le pourcentage d'une fréquence.

Le principal avantage du pourcentage cumulé par rapport à la fréquence cumulée comme mesure de distribution de fréquences est qu'il permet de comparer plus facilement différents ensembles de données.

Les diagrammes de fréquence cumulée et de pourcentage cumulé sont exactement les mêmes, à l'exception de l'échelle de l'axe vertical. En fait, on peut tracer les deux axes verticaux (l'un pour la fréquence cumulée et l'autre pour le pourcentage cumulé) sur le même diagramme.

On calcule le pourcentage cumulé 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 :

pourcentage cumulé = (fréquence cumulée ÷ n) x 100

Exemple 1 – Calcul du pourcentage cumulé

Voici l'épaisseur de la neige enregistrée (mesurée au centimètre près) à Whistler Mountain, en Colombie-Britannique, durant une période de 25 jours :

242, 228, 217, 209, 253, 239, 266, 242, 251, 240, 223, 219, 246, 260, 258, 225, 234, 230, 249, 245, 254, 243, 235, 231, 257.

1. Utilisez les données ci-dessus (soit les mêmes que celles utilisées dans l'exemple2 de la section précédente portant sur la fréquence cumulée) pour :
   - construire un autre tableau de distribution de fréquences;
   - calculer la fréquence de chaque intervalle;
   - déterminer l'extrémité de chaque intervalle;
   - calculer la fréquence cumulée et le pourcentage;
   - enregistrer les résultats dans le tableau.
2. Dessinez un diagramme comportant deux axes verticaux (tracés de chaque côté ou en juxtaposition) : un pour la fréquence cumulée et l'autre pour le pourcentage cumulé. Assurez-vous d'identifier l'axe vertical (axe des y) au moyen des étiquettes fréquence cumulée et pourcentage cumulé d'un côté et de l'autre de cet axe. Identifiez aussi l'axe horizontal (axe des x) à partir de l'autre variable (épaisseur de la neige).

Réponses

1. Les mesures de l'épaisseur de la neige varient entre 209 cm et 266 cm. Pour produire le tableau, il est préférable de regrouper les données par intervalles de classe de 10 cm chacun.

   Chaque intervalle de classe de 10 cm qui se situe entre 200 cm à 270 cm est indiqué dans la colonne Épaisseur de la neige.

   La colonne Fréquence sert à enregistrer le nombre d'observations dans un intervalle donné. Les observations indiquées dans la colonne Comptage ne sont représentées que sous forme numérique.

   Chaque valeur apparaissant dans la colonne Extrémité est le nombre le plus élevé que l'on trouve dans chaque intervalle de classe. Dans l'intervalle de 200 cm à 210 cm, la valeur de l'extrémité est 210.

   La colonne Fréquence cumulée sert à indiquer le total de chaque fréquence qui vient s'ajouter à la dernière enregistrée, comme on l'a vu dans les exercices de la section précédente.

   La colonne Pourcentage cumulé montre la fréquence cumulée, divisée par le nombre total d'observations (25, dans ce cas). On multiplie ensuite le résultat par 100. Ce calcul donne le pourcentage cumulé de chaque intervalle.

Tableau 1. Épaisseur de la neige enregistrée à Whistler Mountain, en Colombie-Britannique, durant une période de25jours

Épaisseur de la neige (x)	Comptage	Fréquence (f)	Extrémité	Fréquence cumulée	Pourcentage cumulé
			200	0	0 ÷ 25 x 100 = 0
200 à < 210	I	1	210	1	1 ÷ 25 x 100 = 4
210 à < 220	II	2	220	3	3 ÷ 25 x 100 = 12
220 à < 230	III	3	230	6	6 ÷ 25 x 100 = 24
230 à < 240	𝍸	5	240	11	11 ÷ 25 x 100 = 44
240 à < 250	𝍸 II	7	250	18	18 ÷ 25 x 100 = 72
250 à < 260	𝍸	5	260	23	23 ÷ 25 x 100 = 92
260 à < 270	II	2	270	25	25 ÷ 25 x 100 = 100

2. Le diagramme doit être identique à celui dessiné dans l'exemple2 de la section sur la fréquence cumulée, à l'exception de l'axe supplémentaire représentant le pourcentage cumulé.

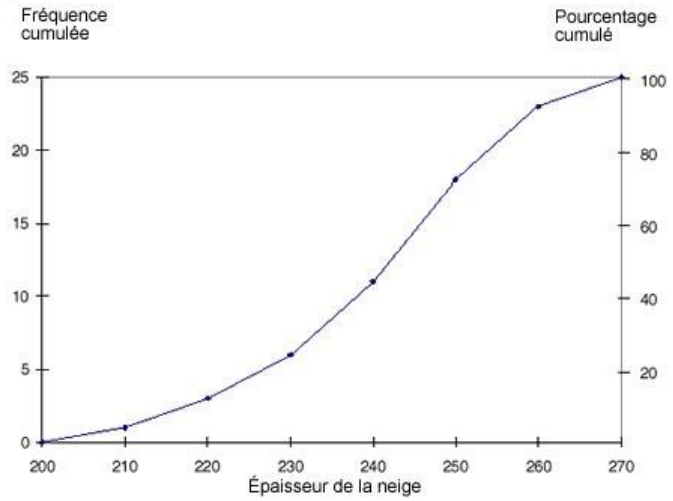
   L'axe du pourcentage cumulé est divisé en cinq intervalles de 20, tandis que celui de la fréquence cumulée est divisé en cinq intervalles de cinq. L'axe de l'épaisseur de la neige est divisé par les extrémités de chaque intervalle de classe de 10 cm.

   En utilisant chaque extrémité pour tracer le diagramme, vous constaterez que la fréquence cumulée et le pourcentage cumulé se trouvent au même point. En utilisant 260 comme extrémité, par exemple, tracez votre point (la fréquence cumulée) au 23e jour. Ce point doit se trouver au même endroit que celui où vous tracerez le pourcentage cumulé (92 %).

   Vous devez porter beaucoup d'attention lorsque vous dessinez un diagramme composé de deux axes y. Par exemple, si vous avez 47 observations à représenter, vous pourriez être tenté d'utiliser des intervalles de cinq, puis de fixer à 50 l'extrémité de l'axe des y de la fréquence cumulée. Si vous procédez ainsi, vous devez vous assurer lorsque vous tracez l'axe des y des pourcentages cumulés d'indiquer l'intervalle de 100 % au même point que la marque 47 inscrite sur l'autre axe des y, et non à celui de la marque 50. Ceci s'explique par le fait que la fréquence cumulée de 47 équivaut à 100 % de vos données. Si vous fixez 100 % comme pourcentage à l'extrémité supérieure de l'échelle, c'est-à-dire à l'endroit où l'intervalle de 50 est représenté, la ligne de fréquence cumulée ne concordera pas avec celle du pourcentage cumulé.

   Les points tracés se joignent pour former une ogive qui, dans la plupart des cas, ressemble à un S étiré. On utilise des ogives pour déterminer le nombre, ou le pourcentage, d'observations qui se situent au-dessus ou au-dessous d'une valeur indiquée. D'après le tableau et le diagramme, par exemple, l'épaisseur de la neige enregistrée durant la période de 25 jours était inférieure à la marque de 260 cm, 92 % du temps.

Figure 1. Épaisseur de la neige enregistrée à Whistler Mountain, en Colombie Mountain, en durant une période de 25 jours



Voici l'information que l'on peut tirer du diagramme ou du tableau :

- durant la période de 25 jours, l'épaisseur de la neige enregistrée était inférieure à la marque de 230 cm, 24 % du temps;
- on dénombre sept jours durant cette période où l'épaisseur de la couche de neige atteignait au moins 250 cm.

Exercices

1. L'ensemble de données qui suit fournit la durée du règne (en années) de divers rois et de diverses reines d'Angleterre depuis la bataille de Hastings, en 1066. Si la durée du règne d'un monarque a été de moins de six mois, elle a alors été arrondie à zéro. Dans le cas contraire, elle a été arrondie à l'année la plus près.

Monarques

Monarque	Années	Monarque	Années
Guillaume le Conquérant	24	Édouard VI	7
Guillaume II	13	Jeanne	0
Henri Ier	34	Marie 1er	5
Étienne	18	Élisabeth 1er	44
Mathilde	1	Jacques 1er	22
Henri II	34	Charles 1er	24
Richard Ier	10	Charles II	25
Jean	18	Jacques II	3
Henri III	56	Guillaume III	13
Édouard 1er	35	Marie II	5
Édouard II	20	Anne	12
Édouard III	50	George 1er	13
Richard II	22	George II	33
Henri IV	14	George III	59
Henri V	9	George IV	10
Henri VI	39	Guillaume IV	7
Édouard IV	22	Victoria	64
Édouard V	0	Édouard VII	9
Richard III	2	George V	26
Henri VII	23	Édouard VII	1
Henri VIII	38	George VI	15

   a. Présentez les données sous la forme d'un diagramme à tiges et à feuilles ordonné.

   b. Y a-t-il des valeurs aberrantes? Si oui, pourquoi?

   c. Décrivez les caractéristiques suivantes de la distribution :

      - le nombre de pointes;
      - sa forme générale;
      - la valeur approximative au centre de la distribution.

   d. Calculez la fréquence cumulée et le pourcentage cumulé.

   e. Dessinez l'ogive avec deux axes verticaux différents, un pour la fréquence cumulée et l'autre pour le pourcentage cumulé.

   f. À l'aide de l'information déjà fournie pour l'exercice, répondez aux questions suivantes :

      - Combien de monarques ont régné moins de 10 ans?
      - Combien de monarques ont régné 50 ans ou plus?

   g. Le monarque anglais actuel est la Reine Élisabeth II. Elle règne depuis le 6 février 1952. Elle a été exclue de l'ensemble original de données. Calculez la durée de son règne et formulez un bref commentaire à son sujet par rapport aux autres souverains.

2. Eliza achète souvent un petit sac de frites au casse-croûte local. Elle était curieuse de savoir si elle en recevait pour son argent et si elle obtenait toujours la même quantité de frites dans chaque sac. Elle a donc décidé de compter et d'enregistrer pendant trois mois les frites incluses dans chacun des sacs. À la fin du troisième mois, elle s'est rendue compte qu'elle avait acheté des frites 30 fois. Voici les données découlant de ses 30 achats de frites au casse-croûte local :

   44, 46, 54, 38, 49, 46, 45, 31, 55, 37, 42, 43, 47, 51, 48 40, 59, 35, 47, 21,43, 37, 45, 38, 40, 32, 50, 34, 43, 54

   a. Présentez les données à l'intérieur d'un diagramme à tiges et à feuilles ordonné. Divisez les tiges, si nécessaire.

   b. Y a-t-il des valeurs aberrantes? Si oui, pourquoi?

c. Décrivez les caractéristiques suivantes de la distribution :
   - le nombre de pointes;
   - la forme générale;
   - la valeur approximative au centre de la distribution.

d. Calculez la fréquence cumulée et le pourcentage cumulé.

e. Dessinez l'ogive avec deux axes verticaux différents, un pour la fréquence cumulée et l'autre pour le pourcentage cumulé.

f. À l'aide de l'information déjà fournie pour l'exercice, répondez aux questions suivantes :
   - Combien de sacs contenaient moins de 40 frites?
   - Quel pourcentage de sacs renfermait 45 frites ou plus?

g. Le casse-croûte local a décidé de créer un slogan pour sa nouvelle campagne de publicité. Complétez l'énoncé suivant à l'aide de l'information tirée des calculs d'Eliza.

   « Nos sacs sont peut-être petits, mais la moitié renferme au moins _____frites »

3. Imaginez que le tableau ci-dessous représente (par groupe d'âge) le nombre de femmes en chômage à la recherche d'un emploi à temps plein.

Tableau 1. Nombre de femmes en chômage à la recherche d'un emploi, par groupe d'âge

Groupe d'âge *	Nombre de femmes
15 à 24	339
25 à 34	273
35 à 44	147
45 à 54	121
55 à 64	22

* On fait la collecte des données sur l'âge selon le nombre d'années cumulées. Ainsi, l'extrémité supérieure de l'intervalle 15 à 24 est 25 (reportez-vous aux diagrammes à tiges et à feuilles figurant dans le chapitre sur l'organisation des données).

a. Les variables sont-elles discrètes ou continues?

b. Copiez le tableau 1 dans votre cahier et calculez la fréquence cumulée et le pourcentage cumulé.

c. Dessinez l'ogive avec deux axes verticaux différents, un pour la fréquence cumulée et l'autre pour le pourcentage cumulé.

d. Pourquoi n'y a-t-il pas de données pour les femmes de moins de 15 ans?

e. Dans quel groupe d'âge se situe la valeur du pourcentage cumulé de 50 %?

f. Quel est le pourcentage de femmes en chômage de moins de 25 ans à la recherche d'un emploi?

g. Quel est le pourcentage de femmes en chômage de 55 ans ou plus à la recherche d'un emploi?

h. Comment le gouvernement canadien utiliserait-il ce genre d'information?

4. Imaginez qu'on a mené une enquête afin de déterminer combien de temps il faut aux employés de Statistique Canada pour se rendre à leur travail et en revenir. Voici les résultats enregistrés, arrondis à la minute, à la suite de cette enquête :

33, 63, 49, 65, 56, 45, 52, 63, 38, 66, 43, 98, 60, 58, 68, 29, 59, 87, 22, 64, 73, 56, 71, 67, 44, 31, 83, 50, 75, 65, 60, 51, 89, 69, 41, 76,58, 62, 25, 52, 64, 77, 61, 55, 80, 45, 12, 69, 40, 37

a. Ces variables sont-elles discrètes ou continues?

b. Présentez les données à l'intérieur d'un tableau de fréquences, à l'aide d'intervalles appropriés, incluant des fréquences relatives et des fréquences en pourcentage. _____

c. Dessinez un histogramme pour représenter les données et tracez le polygone de fréquences.

d. Tracez un diagramme à tiges et à feuilles ordonné pour les données. Y a-t-il des valeurs aberrantes? Si oui, pourquoi?

e. Décrivez les caractéristiques suivantes de la distribution :
   - le nombre de pointes;
   - sa forme générale;
   - la valeur approximative au centre de la distribution.

f. Calculez la fréquence cumulée et le pourcentage cumulé. Trouvez les extrémités et enregistrez les résultats à l'intérieur d'un tableau.

g. Dessinez l'ogive avec deux axes verticaux différents, un pour la fréquence cumulée et l'autre pour le pourcentage cumulé.

h. À l'aide de l'information déjà fournie pour l'exercice, répondez aux questions suivantes :
   - Quel était l'intervalle de temps le plus fréquent qu'il fallait aux employés de Statistique Canada pour se rendre à leur travail et en revenir?
   - Quel pourcentage d'employés a pris plus de 90 minutes pour se rendre au travail et en revenir?
   - Combien d'employés ont pris moins de 40 minutes pour se rendre à leur travail et en revenir?

Activité en classe

Sondez les professeurs de votre école pour déterminer depuis combien de temps ils enseignent (en arrondissant à l'année).

1. Les variables sont-elles discrètes ou continues?
2. Présentez les données à l'intérieur d'un tableau de fréquences, à l'aide d'intervalles appropriés, incluant des fréquences relatives et en pourcentage.
3. À l'aide de l'information déjà fournie pour l'exercice, répondez aux questions suivantes :
    a. Depuis combien d'années la plupart des professeurs enseignent-ils?
    b. Quelle est la différence en pourcentage entre les premier et deuxième nombres les plus fréquents d'années d'enseignement?
4. Dessinez un histogramme pour représenter les données et tracez le polygone de fréquences. Dessinez un diagramme à tiges et à feuilles ordonné pour les données.
5. Y a-t-il des valeurs aberrantes? Si oui, pourquoi?
6. Décrivez les caractéristiques suivantes de la distribution :
    - le nombre de pointes;
    - sa forme générale;
    - la valeur approximative au centre de la distribution.
7. Calculez la fréquence cumulée et le pourcentage cumulé.
8. Dessinez l'ogive avec deux axes verticaux différents, un pour la fréquence cumulée et l'autre pour le pourcentage cumulé.
9. À l'aide des nouvelles données, répondez aux questions suivantes :
    a. Combien de professeurs enseignent depuis plus de 10 ans?
    b. Quel pourcentage de professeurs enseigne depuis plus de 10 ans?
    c. Quel pourcentage de professeurs enseigne depuis moins de 10 ans?
    d. Depuis combien d'années la moitié des professeurs enseignent-ils?
10. Présentez votre analyse dans un rapport renfermant les tableaux et les diagrammes de fréquences et de pourcentages cumulés nécessaires.

Réponses

1.

a.

Tableau 1. Durée du règne (en années) des monarques anglais

Tige	Feuille	Fréquence(f)	Valeur supérieure	Fréquence cumulée	Pourcentage cumulé
0	0 0 1 1 2 3 5 5 7 7 9 9	12	9	12	28,57
1	0 0 2 3 3 3 4 5 8 8	10	18	22	52,23
2	0 2 2 2 3 4 4 5 6	9	26	31	73,80
3	3 4 4 5 8 9	6	38	37	88,09
4	4	1	44	38	90,47
5	0 6 9	3	59	41	97,61
6	4	1	64	42	100,00

b. Les valeurs aberrantes possibles sont 56, 59 et 64. Toutefois, les données présentées ici reposent sur des faits; nous disposons donc des nombres, parce que chacun de ces monarques a accédé au trône dans sa jeunesse et a vécu longtemps. Ce raisonnement sous-entend qu'il n'y a aucune valeur aberrante.

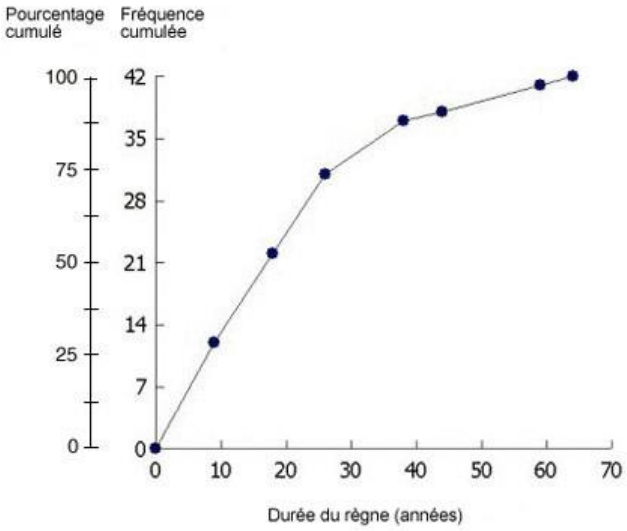
c.

- Une pointe apparaît au début de la distribution.
- La forme générale de la distribution est désaxée vers la droite.
- La valeur approximative au centre de la distribution est 18 ans.

d. Même réponse du Tableau 1

e.

Figure 1. Durée du règne (en années) des monarques anglais



f.

- Le nombre de monarques ayant régné moins de 10 ans s'établit à 12.
- Le nombre de monarques ayant régné 50 ans ou plus est de quatre.

g. La Reine Élisabeth II a célébré le 58e anniversaire de son règne le 6 février 2010. La durée de son règne est déjà bien supérieure au centre de la distribution et deux autres monarques seulement ont régné plus longtemps qu'elle.

2. a.

Tableau 2. Nombre de frites par petit sac

Tige	Feuille
2(0)	1
2(5)	
3(0)	1 2 4
3(5)	5 7 7 8 8
4(0)	0 0 2 3 3 3 4
4(5)	5 5 6 6 7 7 8 9
5(0)	0 1 4 4
5(5)	

b. La valeur aberrante est 21. Cela s'explique peut-être par le fait qu'il n'y avait que 21 frites dans le sac ce jour-là, puisque c'est tout ce qui restait. Il se peut aussi que le nombre consigné soit inexact (c.-à-d., 21 au lieu de 41).

c.

- Le diagramme est unimodal, ce qui signifie que sa distribution ne comporte qu'une seule pointe.
- Si l'on soustrait la valeur aberrante, la forme générale de la distribution est en gros symétrique.
- La valeur approximative au centre de la distribution se situe entre 43 à 44 frites ou 43,5 frites.

d.

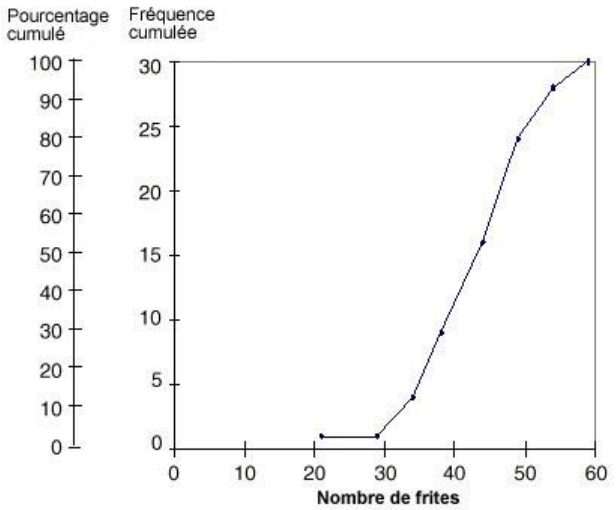
Tableau 3. Nombre de frites par petit sac

Nombre de frites	Fréquence (f)	Valeur supérieure	Fréquence cumulée	Pourcentage cumulé
20 à 24	1	21	1	3,3
25 à 29	0	29	1	3,3
30 à 34	3	34	4	13,3
35 à 39	5	38	9	30,0
40 à 44	7	44	16	53,3
45 à 49	8	49	24	80,0
50 à 54	4	54	28	93,3
55 à 59	2	59	30	100,0
Total	**30**			**100,0**

Nota : Si une des intervalles de classe ne comporte aucun sac, vous devriez toujours utiliser la valeur supérieure comme extrémité. Dans l'exemple ci-dessus, l'intervalle 20 à 24 contient un sac alors que celle de 25 à 29 n'en contient aucune. Pour déterminer la valeur supérieure de l'intervalle 25 à 29, il faut utiliser 29 comme extrémité.

e.

Figure 2. Nombre de frites par petit sac



f.

- Seulement 9 des 30 sacs renfermaient moins de 40 frites.
- La proportion des sacs qui renfermaient 45 frites ou plus est 46,7 %.

g. Le slogan publicitaire devrait se lire comme suit : « Nos sacs sont peut-être petits, mais la moitié renferme au moins 44 frites! »

3.

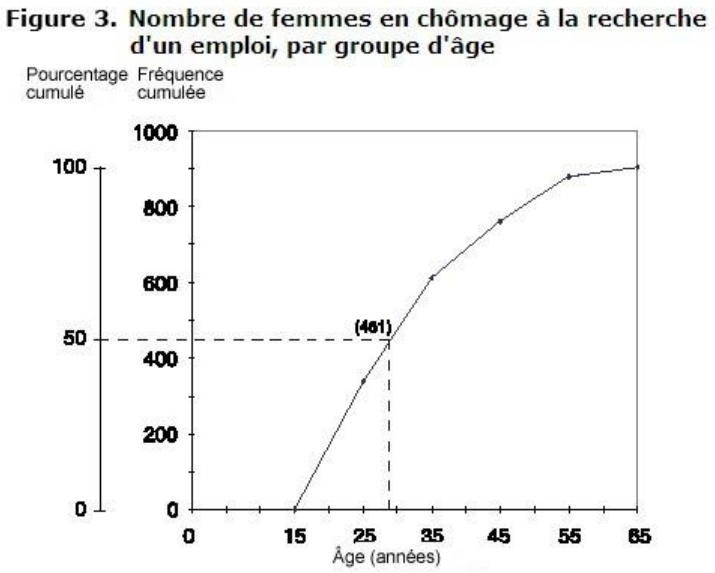
a. Ce sont des variables continues.

b.

Tableau 4. Nombre de femmes en chômage à la recherche d'un emploi, par groupe d'âge

Groupe d'âge	Nombre de femmes	Extrémité	Fréquence cumulée	Pourcentage cumulé
0 à 14	0	15	0	0,0
15 à 24	339	25	339	37,6
25 à 34	273	35	612	67,8
35 à 44	147	45	759	84,1
45 à 54	121	55	880	97,6
55 à 64	22	65	902	100,0

c.



d. Il n'y a pas de données pour les femmes de moins de 15 ans, parce qu'on ne peut pas classer les personnes de cet âge comme chômeurs.

e. Le pourcentage cumulé de 50 % s'inscrit à l'intérieur du groupe d'âge des 25 à 34 ans (il correspond approximativement à 29 ans).

f. La proportion de femmes en chômage de moins de 25 ans et à la recherche d'un travail à temps plein est de 37,6 %.

g. La proportion de femmes en chômage de 55 ans et plus et à la recherche d'un travail à temps plein est de 2,4 %.

h. Le gouvernement canadien pourrait utiliser ces informations pour établir des programmes de création d'emplois axés sur des groupes d'âge particuliers, dans ce cas-ci, sur les moins de 25 ans.

4.

a. Ce sont des variables continues.

b.

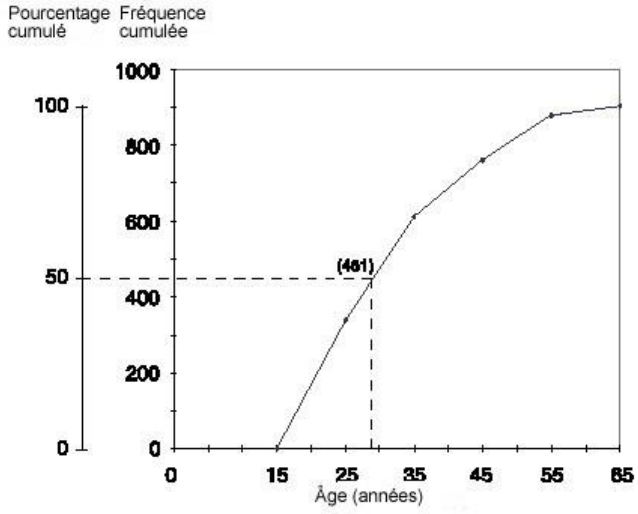
Tableau 5. Temps pris pour se rendre au travail et en revenir par les employés de Statistique Canada, à Ottawa

Temps (x)	Comptage	Fréquence (f)	Fréquence relative	Pourcentage relatif
0 à < 10		0	0,00	0
10 à < 20	𝍸	1	0,02	2
20 à < 30	𝍸	3	0,06	6
30 à < 40	𝍸	4	0,08	8
40 à < 50	𝍸	7	0,14	14
50 à < 60	𝍸 𝍸	10	0,20	20
60 à < 70	𝍸 𝍸 𝍸	15	0,30	30
70 à < 80	𝍸	5	0,10	10
80 à < 90	𝍸	4	0,08	8
90 à < 100	𝍸	1	0,02	2

Total		**50**	**1,00**	**100**

c.

Figure 4. Temps pris pour se rendre au travail et en revenir par les employés de Statistique Canada, à Ottawa



d.

Tableau 6. Temps pris pour se rendre au travail en revenir par les employés de Statistique Canada, à Ottawa

Tige	Feuille
0	
1	2
2	2 5 9
3	1 3 7 8
4	0 1 3 4 5 5 9
5	0 1 2 2 5 6 6 8 8 9
6	0 0 1 2 3 3 4 4 5 5 6 7 8 9 9
7	1 3 5 6 7
8	0 3 7 9
9	8

Il se peut que « 98 » soit une valeur aberrante. Cette valeur pourrait s'expliquer par le fait que la personne en question avait de la difficulté à se rendre au travail ou simplement qu'elle habitait plus loin que la plupart des autres employés.

e.

- Le diagramme est unimodal, ce qui signifie que sa distribution ne comporte qu'une seule pointe.
- La forme générale au centre de la distribution est assez symétrique.
- La valeur approximative au centre de la distribution se situe entre 59 et 60 ou 59,5 minutes.

f.

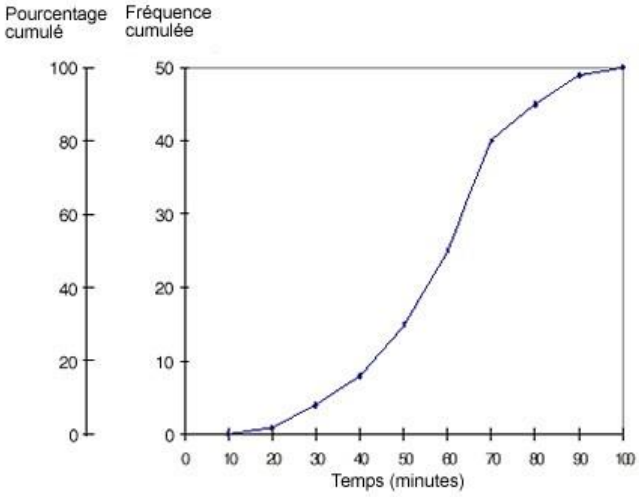
Tableau 7. Temps pris pour se rendre au travail et en revenir par les employés de Statistique Canada, à Ottawa

Temps (x)	Fréquence (f)	Extrémité	Fréquence cumulée	Pourcentage cumulé
0 à < 10	0	10	0	0
10 à < 20	1	20	1	2
20 à < 30	3	30	4	8
30 à < 40	4	40	8	16
40 à < 50	7	50	15	30
50 à < 60	10	60	25	50
60 à < 70	15	70	40	80
70 à < 80	5	80	45	90
80 à < 90	4	90	49	98

90 à < 100	1	100	50	100

g.

Figure 5. Temps pris pour se rendre au travail et en revenir par les employés de Statistique Canada, à Ottawa



h.

- L'intervalle de temps le plus fréquent pris pour se rendre au travail et en revenir par les employés de Statistique Canada est 60 à <70 minutes.
- Seulement 2 % des employés prenaient plus de 90 minutes pour se rendre à leur travail et en revenir.
- Seulement 8 des 50 employés sondés prenaient moins de 40 minutes pour se rendre à leur travail et en revenir.

Mesures de tendance centrale

La meilleure façon de réduire un ensemble de données tout en conservant l'information qu'il contient consiste à résumer l'ensemble par une seule valeur.

Les mesures de tendance centrale (la moyenne, la médiane et le mode) vous aident à saisir, à l'aide d'un nombre unique, ce qui est typique d'un ensemble de données.

La moyenne est la valeur moyenne de toutes les données de l'ensemble.

La médiane est la valeur du milieu lorsque les données de l'ensemble ont été classées en ordre numérique : exactement la moitié des données se situant au-dessus de la médiane et la moitié se situant au-dessous.

Le mode est la valeur la plus fréquente de l'ensemble.

Dans une distribution normale, la moyenne, la médiane et le mode ont des valeurs identiques.

Calcul de la moyenne

On calcule la moyenne d'une variable numérique en additionnant les valeurs de toutes les observations incluses dans un ensemble de données, puis en divisant cette somme par le nombre d'observations qui font partie de l'ensemble. Ce calcul permet d'obtenir la valeur moyenne de toutes les données.

Moyenne = Somme de toutes les valeurs d'observation ÷ nombre d'observations

Il y a deux types de variables : discrètes et continues. On définit les variables discrètes comme des variables qui ne peuvent être divisées intérieurement. Un joueur de hockey, par exemple, peut compter 1 ou 2 buts, mais jamais 1 but et demi. On peut cependant diviser des variables continues en unités plus petites. Un élève peut avoir 11 ans, 7 mois et 3 jours, plutôt que simplement 11 ou 12 ans.

Il est important que vous compreniez la différence entre ces deux types de variables pour pouvoir calculer correctement la moyenne dans une situation donnée. Nous utilisons dans les exemples suivants des variables discrètes pour calculer la moyenne.

- Exemple1–TournoidesoccerauMontRivalI(variablesdiscrètes) Exemple2–Accidentsmortelsdelaroute
- (variablesdiscrètes) Exemple3–TournoidesoccerauMontRivalII(tableauxdefréquence–variablesdiscrètes)
- Exemple4–Taillesde50fillesde10eannée(variablesgroupées–variablesdiscrètesoucontinues) Résumé

Exemple 1 – Tournoi de soccer au Mont Rival I

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?

Moyenne = somme de toutes les valeurs observées ÷ nombre d'observations
= (7 + 5 + 0 + 7 + 8 + 5 + 5 + 4 + 5 + 1) ÷ 10
= 47 ÷ 10
= 4,7

Donc, au cours d'un tournoi de 10 parties, le joueur a compté en moyenne 4,7 buts par partie. La moyenne de 4,7 n'est pas un nombre entier; elle n'a donc de signification que statistiquement. En réalité, il est impossible de compter 4,7 buts, même si on est un marqueur de premier plan.

Voici la notation mathématique pour calculer la moyenne d'une variable discrète :

$$\overline{X} = \frac{\sum x}{n}$$

ou

$\sum$ x ÷ n

où **x** représente une valeur observée, n représente le nombre d'observations incluses dans l'ensemble de données,

$\sum$

x représente la somme de toutes les valeurs x observées,

$\overline{X}$

représente la valeur moyenne de **x**.

Exemple 2 – Accidents de la route

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

Tableau 1. Nombre de décès dus aux accidents de laroute

Année	Décès
1	959
2	1 037
3	960

4	797
5	663
6	652
7	560
8	619
9	623
100	583

Vous le constaterez en utilisant la formule ci-dessous pour calculer la moyenne :

Moyenne =
Σ
x ÷ n

= (959 + 1 037 + 960 + 797 + 663 + 652 + 560 + 619 + 623 + 583) ÷ 10
= 7 453 ÷ 10
= 745,3

Le nombre de personnes tuées annuellement sur les routes s'établit à 745,3.

Pour calculer le taux quotidien moyen de décès attribuables à des accidents de la route, on divise le taux annuel moyen de décès par le nombre de jours dans une année (ici on n'a pas tenu compte des années bissextiles).

= 745,3 ÷ 365
= 2,0

Par conséquent, environ 2 personnes sont mortes en moyenne chaque jour dans des accidents de la route.

Tableaux de fréquences

Un tableau de fréquences indique le nombre d'observations à l'intérieur d'un ensemble de données. On peut l'utiliser aussi bien pour les variables groupées que les variables non groupées.

Par exemple, pour avoir un tableau de fréquences sur l'âge d'une personne dans un ensemble de données, vous pouvez produire un tableau qui comporte des variables (non groupées) relatives à l'âge exact ou des variables par groupe d'âge (variables groupées).

Une variablenongroupée peut être considérée comme une variablegroupée présentant un cas particulier (c.-à-d. un groupe). Vous pouvez calculer la moyenne de la variable discrète à l'aide du tableaudefréquences. Cette méthode permet d'avoir une approximation de la vraie moyenne d'une variable non groupée. Le degré d'exactitude de l'approximation dépend du degré d'égalité de la dispersion des valeurs observées à l'intérieur de chaque groupe.

Exemple 3 –Tournoi de soccer au Mont Rival II

Grouper des observations à l'intérieur de tableaux est utile lorsqu'on traite de grandes quantités de données. On peut présenter à l'intérieur d'un tableau de fréquences les chiffres (tirés de l'exemple ci-dessous) sur les buts comptés par le joueur en tête au tournoi de soccer :

Tableau 2. Fréquence des buts comptés par le joueur en tête au tournoi de soccer de Mont Rival

Nombre de buts (x)	Fréquence (f)	Nombre total de buts comptés (xf)
0	1	0
1	1	1
4	1	4
5	4	20
7	2	14
8	1	8
Total (Σ)	10	47

La notation mathématique change légèrement, parce que les observations sont groupées.

La moyenne d'une variable discrète à l'intérieur d'un tableau de fréquences se calcule comme suit :

$$\bar{X} = \frac{\Sigma xf}{\Sigma f}$$

ou
Σ
xf ÷
Σ
f

- x représente une valeur observée,
- xf représente le résultat d'une valeur observée, multiplié par la fréquence,
- xf représente la somme de toutes les valeurs xf,

- f représente la somme de toutes les fréquences,
- représente la valeur moyenne de x.

La moyenne des buts comptés par le joueur en tête se calcule comme suit :

Moyenne =
Σ
xf ÷
Σ
f

= (0 + 1 + 4 + 20 + 14 + 8) ÷ (1 + 1 + 1 + 4 + 2 + 1)
= 47 ÷ 10
= 4,7

Il s'agit d'une moyenne exacte, puisque la variable est non groupée. L'exemple qui suit montre ce qui se produit lorsqu'on utilise des variables groupées.

Exemple 4 – Tailles de 50 filles de 10e année

Le tableau qui suit montre les tailles de 50 filles de 10e année sélectionnées au hasard. Quelle est la taille moyenne des filles?

Déterminez le point milieu de chaque intervalledeclasse avant de calculer la moyenne à partir d'un tableau de fréquences.

Tableau 3. Taille de 50 filles de 10e année

Taille (cm)	Point milieu (x)	Fréquence (f)	Total des points milieux (xf)
150 –< 155	152,5	4	610,0
155 –< 160	157,5	7	1 102,5
160 –< 165	162,5	18	2 925,0
165 –< 170	167,5	11	1 842,5
170 –< 175	172,5	6	1 035,0
175 –< 180	177,5	4	710,0
-	-	50	8 225,0

Le calcul est le même que celui utilisé dans l'exemple ci-dessus sur le tournoi de soccer, sauf que xf représente maintenant le résultat du point milieu de l'intervalle, multiplié par la fréquence de ce dernier. Cette approximation est nécessaire puisque nous ne connaissons pas la taille exacte de chaque fille.

Par conséquent, nous devons traiter toutes les tailles consignées comme si elles représentaient le point milieu de leur intervalle respectif. Par exemple, comme l'intervalle 150 –< 155cm compte quatre filles, nous devons traiter chacune d'elles comme si elles mesuraient 152,5 cm. Comme il a été mentionné dans l'exemple sur le tournoi de soccer, l'exactitude de l'approximation dépendra de la mesure dans laquelle la taille de chacune des filles se rapprochera du point milieu de leur intervalle respectif.

Ainsi,

Moyenne =
Σ
xf ÷
Σ
f

= (610,0 + 1 102,5 + 2 925,0 + 1 842,5 + 1 035,0 + 710,0) ÷ (4 + 7 + 18 + 11 + 6 + 4)
= 8 225,0 ÷ 50
= 164,5 cm

La taille moyenne des 50 filles de 10e année est donc 164,5 cm.

Résumé

La moyenne sert à calculer d'autres statistiques (comme la variance) et n'existe pas pour des distributions de fréquences groupées qui sont ouvertes. Elle n'est souvent pas appropriée pour des distributions désaxées (ou déséquilibrées) comme de l'information salariale. (Voir Mesuresdedispersion pour plus de renseignements sur la variance.)

Calcul de la médiane

Lorsqu'on ordonne les observations d'une variable, la valeur médiane correspond à l'observation qui se trouve au point milieu de cette liste ordonnée. Elle correspond plus précisément à un pourcentagecumulé de 50 % (c'est-à-dire que 50 % des valeurs sont supérieures à la médiane et 50 % lui sont inférieures). La position de la médiane est :

la valeur à la position (n + 1) ÷ 2, le n désignant le nombre de valeurs dans un ensemble de données.

Pour calculer la médiane, il faut d'abord ordonner les données (les trier dans l'ordre ascendant). La médiane est le nombre qui se situe au point milieu.

Médiane = la valeur du milieu d'un ensemble de données ordonnées

On calcule habituellement la médiane pour des variablesnumériques, mais on peut également la calculer pour des variables catégoriques qui sont séquentielles, comme les catégories d'un sondage sur la satisfaction (excellent, bon, satisfaisant et médiocre). On peut classer ces catégories qualitatives en ordre donc elles sont considérées comme des mesures ordinales.

Données brutes

Dans le cas des donnéesbrutes, la médiane est la valeur pour laquelle exactement la moitié des données se situent au-dessus, tandis que l'autre moitié lui est inférieure. Ces deux moitiés se rencontrent à la position médiane. Si le nombre d'observations est impair, la médiane s'avère parfaite et le nombre à la position de la médiane sera un nombre entier. Par ailleurs, si le nombre d'observations est pair, la position de la médiane sera une décimale. Vous devez trouver le point milieu entre les valeurs d'un côté et de l'autre de la position de la médiane.

Exemple 1 – Données brutes (variables discrètes)

Supposons qu'un champion coureur 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 et 25,0 secondes.

Comment calcule-t-on le temps médian?

On commence d'abord en classant les valeurs dans l'ordre ascendant : 25,0, 25,2, 25,6, 25,7 et 26,1.

On utilise ensuite la formule qui suit pour déterminer quelle valeur constitue la valeur du milieu. Rappelons que n représente le nombre de valeurs qui se trouve dans l'ensemble de données.

Médiane = la valeur à la position (n + 1) ÷ 2
= (5 + 1) ÷ 2
= 3

Donc la troisième valeur dans l'ensemble de données sera la médiane. Comme la troisième valeur est 25,6 secondes, elle constituera le temps médian.

= 25,6 secondes

Exemple 2 – Données brutes (variables discrètes)

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

On place, encore une fois, les données dans l'ordre ascendant : 24,7, 25,0, 25,2, 25,6, 25,7, 26,1. On utilise ensuite la même formule pour calculer le temps médian.

Médiane = la valeur à la position (n + 1) ÷ 2
= (6 + 1) ÷ 2
= 7 ÷ 2
= 3,5

Comme il y a un nombre pair d'observations dans cet ensemble de données, il n'y a pas de valeur du milieu distincte. La médiane tombe entre les 3e et 4e valeurs, c'est-à-dire la $3,5^{e}$ observation dans l'ensemble de données. On calcule donc la médiane en établissant la moyenne des deux valeurs du milieu, 25,2 et 25,6. Utilisez la formule ci-dessous pour calculer la valeur moyenne.

Moyenne = (valeur au-dessous de la médiane + valeur au-dessus de la médiane) ÷ 2
= (troisième valeur + quatrième valeur) ÷ 2
= (25,2 + 25,6) ÷ 2
= 50,8 ÷ 2
= 25,4

La valeur 25,4 se situe directement entre les troisième et quatrième valeurs dans cet ensemble de données. Le temps médian est donc de 25,4 secondes.

Distribution de fréquences non groupées

Pour trouver la médiane à partir des fréquences cumulées (ou le nombre d'observations qui se situent au-dessus ou au-dessous d'une valeur particulière dans un ensemble de données), vous devez calculer la première valeur à l'aide d'une fréquence cumulée égale ou supérieure à la valeur médiane. Si la valeur médiane est exactement supérieure de 0,5 à la fréquence cumulée de la valeur précédente, la médiane constituera alors le point milieu entre les deux intervalles.

Exemple 3 – Tableau de fréquences non groupées (variables discrètes)

Imaginez que le nombre de coups de circuit frappés en dix parties par l'équipe de baseball de votre école s'établit comme suit :

4, 5, 8, 5, 7, 8, 9, 8, 8, 7

Si vous deviez consigner le nombre total de coups de circuit frappés dans un tableau de fréquences, quel serait la médiane?

Vous devez d'abord classer les résultats dans l'ordre ascendant :

4, 5, 5, 7, 7, 8, 8, 8, 8, 9

Construisez ensuite un tableau comportant deux colonnes. L'étiquette de la première colonne devrait s'intituler « Nombre de coups de circuit frappés ». Ensuite, indiquez le nombre de coups de circuits frappés que l'équipe pourrait possiblement réaliser. Vous pouvez commencer par 0 en augmentant progressivement jusqu'à 10, mais comme l'équipe n'a jamais réalisé plus de 4 coups de circuits frappés, vous pourriez peut-être commencer par le nombre 4.

L'étiquette de la deuxième colonne devrait s'intituler « Fréquence ». Inscrivez dans cette colonne le nombre de fois que l'équipe a réalisé 4 coups de circuit frappés, 5 coups de circuit frappés et ainsi de suite. Dans ce cas, bien que l'équipe ait réalisé 4 coups de circuit frappés une seule fois, elle a en réalisé 5 à deux reprises. Si vous additionnez tous les chiffres dans la colonne « Fréquence », le total devrait être 10 (soit le nombre de parties disputées).

Tableau 1. Nombre de coups de circuit frappés en 10 partiesdebaseball

Nombre de coups de circuit frappés (x)	Fréquence (f)
4	1
5	2
6	0
7	2
8	4
9	1

Vous devez utiliser la même formule pour trouver la médiane :

Médiane = la valeur à la position (n + 1) ÷ 2
= (10 + 1) ÷ 2
= 11 ÷ 2
= 5,5
= la médiane est la 5,5^{e} valeur dans l'ensemble de données

Pour obtenir la médiane, vous devez additionner chaque nombre inscrit dans la colonne « Fréquence » jusqu'à ce que vous obteniez 5 comme total (puisque 10 parties ont été disputées, le total des nombres restant dans la colonne sera aussi égal à 5). Vous obtiendrez le total de 5 après avoir additionné toutes les fréquences allant jusqu'à 7 coups de circuit frappés inclusivement. Le prochain ensemble de cinq commence par les fréquences pour 8 coups de circuit frappés. La médiane (c'est-à-dire la 5,5^{e} valeur) se situe entre les cinquième et sixième valeurs. Par conséquent, la médiane se situe entre les 7 et 8 coups de circuit frappés.

Si vous calculez la moyenne de ces valeurs (en utilisant la même formule que celle employée pour l'exemple 2), le résultat doit être 7,5.

Moyenne = (valeur précédent le milieu + valeur qui suit le milieu) ÷ 2
= (cinquième valeur+ sixième valeur) ÷ 2
= (7 + 8) ÷ 2
= 15 ÷ 2
= 7,5

Techniquement, la médiane doit correspondre à une variable possible. Dans l'exemple ci-dessus, les variables sont discrètes et sont toujours des nombres entiers. Par conséquent, 7,5 n'est pas une variable possible, puisque personne ne peut frapper 7,5 coups de circuit. Ce nombre est significatif seulement dans un contexte statistique. Certains mathématiciens pourraient faire valoir que 8 est une médiane plus appropriée.

Distribution de fréquences groupées

Il est parfois préférable de ne pas inscrire toutes les variables individuelles dans un tableau de distribution de fréquences lorsque celui-ci deviendrait trop long et difficile à utiliser. Pour simplifier l'opération, divisez l'étendue de données en intervalles, puis indiquez les intervalles dans un tableau de distribution de fréquences qui comprendra une colonne pour le pourcentage cumulé. (Pour plus de renseignements, consultez la section Fréquence cumulée.)

Le calcul de la médiane prend un peu plus de temps, parce que les données ont été groupées en intervalles, ce qui fait que toute l'information originale a été perdue. Dans certains manuels, on utilise seulement le point milieu d'un intervalle comme médiane. Toutefois, cela est une simplification de la valeur réelle. Effectuez les calculs suivants pour trouver la médiane dans une distribution de fréquences groupées.

1. Déterminez où se trouve la médiane parmi les intervalles à l'aide de la formule **(n + 1) ÷ 2**. Utilisez la valeur qui en résultera, puis additionnez tous les chiffres inscrits dans la colonne des fréquences jusqu'à ce que vous obteniez ce nombre (comme dans l'exemple 3). Si votre médiane, par

exemple, se trouve à la valeur 13,5, additionnez les fréquences jusqu'à ce que vous atteigniez les 13e et 14e valeurs. L'intervalle dans laquelle on retrouvera ces valeurs s'appelle le groupe médian.

2. Déterminez le pourcentage cumulé de l'intervalle qui précède le groupe médian. Désignez cette valeur par la lettre **A**.

3. À l'aide du pourcentage cumulé, calculez combien de nombres sont nécessaires pour atteindre 50 % du pourcentage cumulé total. Intitulez cette valeur au moyen de la lettre B. Utilisez la formule qui suit pour calculer la valeur de **B** :

   B = 50 - A

4. Déterminez l'étendue (c.-à-d., combien il y a de nombres dans l'intervalle). Appelez cette valeur C. Calculez ensuite le pourcentage de l'intervalle médian. Désignez cette valeur par la lettre **D**.

5. Déterminez le nombre de valeurs dans la médiane que vous devez compter pour atteindre 50 % du total de l'ensemble des données à l'aide de la formule qui suit. Désignez cette valeur par la lettre **E**.

   E = (B ÷ D) x C

6. Calculez la médiane en additionnant la valeur de E à celle qui se trouve à l'extrémité de l'intervalle médian :

   Médiane = valeur inférieure + E

   Puisque E = (B ÷ D) x C, cette formule peut aussi être décrite comme suit :

   Médiane = valeur inférieure + (B ÷ D) x C

Si la fréquence cumulée de l'intervalle est exactement 50 %, la valeur médiane constituera donc l'extrémité de cet intervalle.

Voici un exemple pour vous aider à mieux comprendre!

Exemple 4 – Variables groupées – distribution de fréquences (variables continues ou discrètes)

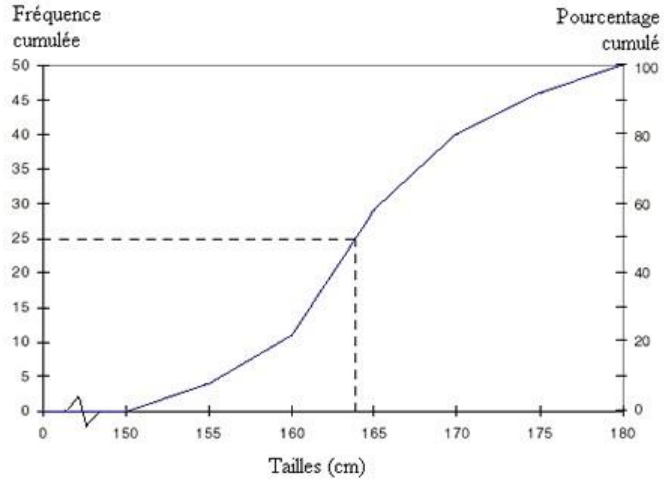
En utilisant les mêmes renseignements que dans l'exemple 4 de la section sur la moyenne, imaginez que vous avez effectué une enquête auprès de 50 filles de 10e année afin de savoir quelle est la taille de chacune d'elles, en centimètres. Après avoir rassemblé toutes vos données, créez un tableau de distribution de fréquences qui ressemble un peu à celui-ci :

Tableau 2. Tailles des filles de 10e année

Taille (cm)	Fréquence (f)	Extrémité (x)	Fréquence cumulée	Pourcentage	Pourcentage cumulé
150 à < 155	4	155	4	8	8
155 à < 160	7	160	11	14	22
160 à < 165	18	165	29	36	58
165 à < 170	11	170	40	22	80
170 à < 175	6	175	46	12	92
175 à < 180	4	180	50	8	100

À l'aide des données groupées, tracez un diagramme de fréquences cumulées pour accompagner votre tableau. Dessinez dans votre diagramme les extrémités des intervalles de tailles ainsi que les nombres se rapportant à la fréquence cumulée et au pourcentage cumulé.

Figure 1. Tailles des filles de 10e année



Essayez de trouver la médiane simplement en observant le diagramme. La médiane est le point où l'axe x (la taille) croise le point milieu (25) de l'axe y (la fréquence cumulée). Vous constaterez que la médiane est environ 164 cm. Si vous faites un calcul mathématique, vous constaterez que la valeur réelle se situe à 163,9 cm. Voici comment :

1. À l'aide de l'information fournie au tableau 2 :

   Médiane = la valeur à la position (n + 1) ÷ 2
   = (50 + 1) ÷ 2
   = 51 ÷ 2
   = 25,5

   En additionnant les fréquences, on constate que la médiane (25,5) se trouve dans le groupe médian de l'intervalle 160 à < 165 cm.

2. Le pourcentage cumulé de l'intervalle précédent (**A**) est 22.

3. Le pourcentage nécessaire pour obtenir 50 % du pourcentage cumulé total (**B**) est 28.

   B = 50 - A
   = 50 - 22
   = 28

4. L'étendue de l'intervalle médian (C) est 5, alors que le pourcentage de l'intervalle médian (**D**) est 36.

5. Le nombre de valeurs que vous devez compter pour obtenir 50 % du total de l'ensemble des données est 3,9.

   E = (B ÷ D) x C
   = (28 ÷ 36) x 5
   = 3,9

6. Comme la valeur la plus faible de l'intervalle médian est 160, vous devez additionner la valeur de **E** pour obtenir la médiane de 163,9 cm.

   Médiane = valeur la plus faible de l'intervalle médian + (B ÷ D) x C
   = 160 + (28 ÷ 36) x 5
   = 160 + 3,9
   = 163,9 cm

Diagrammes à tiges et à feuilles

Les diagrammes à tiges et à feuilles ordonnés simplifient le calcul de la médiane, particulièrement dans le cas où les fréquences cumulées ont déjà été calculées. Examinons les tailles des 50 filles de 10e année à l'aide d'un diagramme à tiges et à feuilles. (Voir le chapitre intitulé Organisation des données pour plus de renseignements sur la façon de créer ces tableaux.)

Exemple 5 – Diagramme à tiges et à feuilles

Tableau3.Taillesdesfillesde10^{e}année

Tige* (cm)	Feuille	Fréquence cumulée
15$^{(0)}$	0 1 1 4	4
15$^{(5)}$	5 6 7 7 8 8 8	11
16$^{(0)}$	0 1 1 1 1 2 2 2 2 2 2 2 3 3 3 4 4 4 4	29
16$^{(5)}$	5 5 5 5 6 6 6 7 7 8 9	40
17$^{(0)}$	0 0 1 2 3 3	46
17$^{(5)}$	6 6 7 8	50

***Nota :** Les tiges ont été divisées en intervalles plus petits. La tige 15(0) signifie que toutes les données s'inscrivent à l'intérieur de l'intervalle 150 à 154. La tige 15(5) signifie que les données se situent à l'intérieur de l'intervalle de 155 à 159.

Comme il y a 50 éléments de données, la valeur de la médiane correspond à la 25,5^{e} observation.

Médiane = la valeur à la position (n +1) ÷ 2
= (50 + 1) ÷ 2
= 51 ÷ 2
= 25,5

La médiane se situe donc entre les 25e et 26e valeurs. Pour connaître quelles sont ces valeurs, comptez chaque valeur inscrite dans la colonne des feuilles jusqu'à ce que vous atteigniez les 25e et 26e valeurs. Ces valeurs se trouvent dans l'intervalle 16(0), c'est-à-dire l'intervalle 160–164. Les nombres dans la colonne des feuilles représentent les nombres qui s'inscrivent dans l'intervalle (p.ex., 3 représente 163). La médiane se situe donc entre les centimètres 163 (25e valeur) et 164 (26e valeur). Il faut calculer la moyenne de ces deux valeurs pour trouver la médiane.

Moyenne = (valeur avant la médiane + valeur après la médiane) ÷ 2
= (25e valeur + 26^{e} valeur) ÷ 2
= (163 + 164) ÷ 2
= 327 ÷ 2
= 163,5

Puisque la taille est une variable continue, la valeur 163,5 cm est considérée comme une médiane acceptable.

La médiane obtenue à partir du diagramme de fréquences cumulés (164 cm) n'est pas la même que celle obtenue à partir du calcul utilisé dans l'exemple 4 (163,9 cm) ou du diagramme à tiges et à feuilles (163,5 cm). Cela s'explique par le fait qu'on peut seulement avoir une approximation de la médiane, à moins que le diagramme ne puisse être dessiné précisément à l'aide de toute l'information utilisée.

Les calculs de l'exemple 4 ne sont que des approximations, puisque des données groupées ne vous permettent pas de savoir la répartition du 36 % des 50 filles qui s'inscrivent dans l'intervalle médian. Par conséquent, on pourrait supposer qu'elles ont été réparties uniformément dans l'intervalle, si bien que la médiane pourrait varier légèrement. Toutefois, un diagramme à tiges et à feuilles est la méthode la plus précise qui soit pour obtenir la médiane, puisqu'on utilise la totalité des valeurs réelles.

Comparaison de la moyenne et de la médiane

Il est possible que la moyenne et la médiane d'une distribution aient la même valeur. C'est toujours le cas si la distribution est symétrique comme dans une distributionnormale. Les deux valeurs seront proches l'une de l'autre si la distribution est en gros symétrique.

Dans l'exemple des tailles de 50 filles de 10e année, la moyenne (164,5 cm) est très proche de la valeur de la médiane (163,5 cm), parce que la distribution est en gros symétrique (voir le diagramme à tiges et à feuilles fourni dans l'exemple ci-dessus).

Toutefois, un chiffre ou un nombre peut modifier la moyenne sans influencer la médiane.

Exemple 6 – Comparaison de la moyenne et de la médiane

Examinons les ensembles de données suivants qui représentent le nombre de buts comptés par 3 joueurs en 11 parties de baseball.

Noémie : 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3
Moyenne = 22 ÷ 11 = 2
Médiane = 2

Jeremy : 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4
Moyenne = 23 ÷ 11 = 2,1
Médiane = 2

Alexandre : 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 14
Moyenne = 33 ÷ 11 = 3
Médiane = 2

Les trois ensembles de données ci-dessus sont identiques, à l'exception des dernières valeurs d'observation (3, 4 et 14).

La médiane ne varie pas, parce qu'elle ne dépend que de la valeur d'observation du milieu. Cependant, la moyenne varie beaucoup parce qu'elle dépend de la valeur moyenne de toutes les observations. Par conséquent, dans l'exemple ci-dessus, lorsque la valeur de la dernière observation augmente, la moyenne en fait autant.

Dans le troisième ensemble de données, la valeur de 14 varie beaucoup de toutes les autres valeurs. Quand une observation est très différente de toutes les autres d'un ensemble de données, on l'appelle une valeur aberrante. (Voir la section sur les diagrammes à tiges et à feuilles pour plus de renseignements sur les valeurs aberrantes.) La moyenne est la mesure de la tendance centrale la plus influencée par les valeurs aberrantes.

Il peut parfois y avoir des valeurs aberrantes en raison d'une erreur ou d'une déformation délibérée de l'information. Si tel est le cas, on devrait exclure les valeurs aberrantes de la mesure de tendance centrale. Dans d'autres cas, les valeurs aberrantes s'avèrent très utiles pour démontrer la mesure dans laquelle une valeur peut différer des autres.

Exemple 7 – Comparaison de la moyenne et de la médiane

Lorsqu'un article de journal renvoie aux prix des maisons, on y cite habituellement le prix médian. Pourquoi utilise-t-on cette mesure, plutôt que la moyenne?

Il y a beaucoup de maisons à prix modérés, mais certaines sont coûteuses et quelques-unes sont très chères. Le prix moyen pourrait être assez élevé, puisqu'il inclut les prix des maisons les plus coûteuses. La médiane donne donc une valeur plus exacte et réaliste des prix auxquels la plupart des gens sont confrontés.

En résumé, la médiane est le nombre au centre d'une distribution. La médiane est utile lorsqu'une distribution est déxaxée (ou déséquilibrée), parce que cette mesure n'est pas du tout influencée par les valeurs aberrantes.

Exemple 8 – Comparaison de la moyenne et de la médiane

Supposons que vous voulez savoir combien d'argent une famille pourrait dépenser pour l'achat d'une maison. Cela dépendrait du revenu total de cette famille.

Pour une famille de cinq personnes (deux parents qui travaillent contre rémunération et trois enfants ne réalisant aucun revenu), le revenu moyen de chaque membre de cette famille est le revenu total de cette dernière divisé par cinq (p.ex., 60 000 $ ÷ 5 = 12 000 $). Toutefois, le revenu médian de la famille en question serait zéro, parce que plus de la moitié de ses membres ne gagnent rien. Dans certains cas, la moyenne peut donc être plus révélatrice que la médiane.

Exemple 9 – Comparaison de la moyenne et de la médiane

Quand vous voulez déterminer si un pays est riche, vous pourriez envisager d'utiliser la médiane comme mesure de tendance centrale, plutôt que la moyenne.

Le revenu familial moyen pourrait être assez élevé si les revenus étaient fortement concentrés dans peu ou très peu de familles à l'aise (en dépit du fait que la plupart des familles ne gagneraient essentiellement rien). Le revenu familial médian serait donc une mesure plus significative (la moitié des familles gagneraient moins que le revenu médian et au moins 50 % gagneraient autant ou plus que le revenu médian).

Exemple 10 – Comparaison de la moyenne et de la médiane

Supposons que vous postulez un emploi de comptable dans plusieurs grandes entreprises. Vous voulez vous faire une idée du montant d'argent que vous pourriez gagner en cinq ans en joignant les rangs de l'une ou l'autre des entreprises. Vous pourriez examiner les salaires des comptables de chaque entreprise cinq ans après leur engagement.

Un seul salaire très élevé pourrait accroître le salaire moyen, ce qui risquerait de ne pas refléter le salaire typique. D'un autre côté, la moitié des comptables gagnent le salaire médian ou moins et l'autre moitié, le salaire médian ou plus. La mesure de tendance centrale qui vous donnerait une meilleure idée d'un salaire typique serait donc la médiane.

Exemple 11 – Comparaison de la moyenne et de la médiane

En choisissant une mesure de tendance centrale qui vous est favorable, vous pouvez tromper des gens à l'aide de statistiques. En fait, cela se fait couramment.

Imaginez que vous êtes le propriétaire d'une boulangerie se spécialisant dans la fabrication et la vente de gâteaux de fête et d'énormes gâteaux de noces.

Il serait peut-être dans votre intérêt d'affirmer à vos clients que les prix de vos produits ont été réduits et à vos actionnaires que ces prix ont été augmentés. Supposons que l'an dernier vous avez vendu 100 000 gâteaux de fête 10 $ chacun et 1 000 gâteaux de noces 1 000 $ chacun. Cette année, vous avez vendu 100 000 gâteaux de fête 8 $ chacun et 1 000 gâteaux de noces 1 200 $ chacun.

- Le prix médian des 101 000 gâteaux vendus l'an dernier est 10 $, parce que plus de 50 % de ceux-ci étaient des gâteaux de fête. Le prix médian des 101 000 gâteaux vendus cette année est 8 $.
- Le prix moyen des 101 000 gâteaux vendus l'an dernier est 19,80 $.

  (100 000 x 10 $ + 1 000 x 1 000 $) ÷ 101 000 = 19,80 $
- Le prix moyen des 101 000 gâteaux vendus cette année est également 19,80 $.

  (100 000 x 8 $ + 1 000 x 1 200 $) ÷ 101 000 = 19,80 $

Le prix moyen par gâteau vendu est le même pour les deux années. Vos recettes totales et le nombre d'articles vendus étaient aussi les mêmes. Vous pouvez faire en sorte que les données semblent indiquer des résultats contradictoires en choisissant la mesure de tendance centrale appropriée.

Il est important de souligner que vous n'êtes pas tenu de n'utiliser qu'une seule mesure de la tendance centrale. De façon à pouvoir obtenir le maximum d'information sur les données, vous pouvez utiliser à la fois la moyenne et la médiane.

Calcul du mode

Le mode est la valeur la plus souvent observée dans un ensemble de données. Il se peut qu'il n'y ait dans cet ensemble aucun mode si aucune des valeurs n'y apparaît plus qu'une autre. Il est aussi possible qu'il y ait dans le même ensemble deux (bimodal), trois (trimodal) ou quatre modes ou plus (multimodal). Dans le cas des distributions de fréquences groupées, la classe modale est celle dont la fréquence est la plus élevée.

Mode = la donnée la plus fréquemment observée

Un ensemble de données peut compter plus d'un mode. Le mode n'indique pas nécessairement le centre d'un ensemble de données. Il se trouvera à proximité de la moyenne et de la médiane si la distribution des données est normale ou quasi-normale. En fait, si la distribution est la fois symétrique et unimodale, il se peut que la moyenne, la médiane et le mode aient tous la même valeur.

Variables nominales ou discrètes

Pour les variables nominales ou discrètes, le mode est simplement la valeur la plus observée. Pour déterminer le mode, il n'est pas nécessaire d'ordonner les observations, même s'il est recommandé de le faire pour faciliter le calcul.

Exemple 1 – Variables nominales ou discrètes

Lors d'un tournoi d'hockey, Audrey a compté 7, 5, 0, 7, 8, 5, 5, 4, 1 et 5 buts en dix parties. Le mode de son ensemble de données est 5, puisque cette valeur apparaît le plus souvent (quatre fois). Il y a fort à parier que Audrey compterait 5 buts si l'on choisissait une partie au hasard.

Exemple 2 – Variables nominales ou discrètes

En 12 parties de basket-ball, Grabriel a compté 14, 14, 15, 16, 14, 16, 16, 18, 14, 16, 16 et 14 points. Cet ensemble de données est bimodale, 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 cinq fois).

Exemple 3 – Variables nominales ou discrètes

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

Classez vos données dans l'ordre suivant :

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

Trouvez et comparez la moyenne, la médiane et le mode.

Mode = la valeur la plus souvent observée = 0

Le mode est 0, puisqu'il s'agit de la valeur la plus souvent observée. Si l'on choisissait au hasard une partie, le mode nous dirait qu'il y a fort à parier que Jérôme ne marquera aucun touché.

Moyenne =

Σ

x ÷ n
= 14 ÷ 14
= 1

Toutefois, Jérôme marquera en moyenne un touché par partie, même si le mode indique qu'il n'a marqué aucun touché dans un grand nombre de matchs. Dans ce cas, le mode ne fournit pas une mesure utile de son rendement.

Médiane = la valeur (n + 1) ÷ 2
= (14 + 1) ÷ 2
= 15 ÷ 2
= 7,5

Moyenne = (valeur en dessous de la médiane + valeur au-dessus de la médiane) ÷ 2
= (septième + huitième valeurs) ÷ 2
= (1 + 1) ÷ 2
= 1

Puisque le nombre de valeurs dans l'ensemble de données est pair, la médiane ne s'inscrit pas exactement dans le centre de l'ensemble des données. Il a donc été nécessaire de calculer la médiane à l'aide des équations ci-dessus. D'après les résultats obtenus, la médiane indique que Jérôme ne marquera qu'un touché par partie.

Variables groupées (continues ou discrètes)

Lorsqu'on groupe des variables continues ou discrètes à l'intérieur de tableaux, on définit le mode comme étant l'intervalle de classe à l'intérieur duquel se situent la plupart des observations. On l'appelle l'intervalle de classe modale.

Dans l'exemple des tailles des 50 filles de 10e année, l'intervalle de classe modale serait 160 à 165 cm, puisqu'il s'agit de l'intervalle qui compte le plus grand nombre d'observations.

On utilise rarement le mode comme mesure de tendance centrale pour des variables numériques. Pour des variables nominales, cependant, le mode est plus utile, parce que la moyenne et la médiane n'ont aucun sens.

Vous pourriez ensuite déterminer le milieu de l'étendue de la classe modale. Le milieu de l'étendue est simplement le point milieu entre la valeur la plus élevée et la valeur la plus faible d'une classe modale. On n'utilise pas très souvent le mode de concert avec le milieu de l'étendue, parce qu'il ne donne qu'une estimation très grossière de la moyenne.

Contrairement à la moyenne et à la médiane, le mode peut être utilisé avec des données nominales. Il peut y avoir un mode comme il peut ne pas y en avoir, et il est possible que le mode comporte plus d'une valeur.

Résumé

En règle générale, ce sont les circonstances qui dictent le choix entre la moyenne, la médiane ou le mode comme mesure de tendance centrale. La moyenne tend à être la mesure de la tendance centrale qui convient le plus pour obtenir un total, parce qu'on divise le total par le nombre de données. Par exemple, le revenu moyen des membres d'une famille vous permet de savoir combien chacun d'entre eux dépense sur les nécessités courantes de la vie. La médiane est utile dans le cas des moyennes, alors que le mode sert à décrire le cas le plus typique.

Exercices

1. Pour les ensembles de données suivants, trouvez (à une décimale près) :
    i. la moyenne;
    ii. la médiane;
    iii. le mode.

    a. 0 – 0 – 0 – 0 – 1 – 0 – 0 – 0 – 0 – 0 – 0 Réponse 1a
    b. 2 – 1 – 2 – 3 – 1 – 3 – 0 – 2 – 4 – 2 – 2 Réponse 1b
    c. 2,4 – 3,9 – 1,8 – 1,7 – 4,0 – 2,1 – 3,9 – 1,5 – 3,9 – 2,6 Réponse 1c
    d. 153,8 – 154,7 – 156,9 – 154,3 – 152,3 – 156,1 – 152,3 Réponse 1d

2. Pour les ensembles de données suivants, trouvez :
    i. la moyenne;
    ii. la médiane;
    iii. le mode.

    Décrivez brièvement les positions de la moyenne, de la médiane et du mode et leur relation les unes par rapport aux autres pour chaque ensemble de données.

    a.

Tableau 1

x	fréquence
-2	3
-1	7
0	8
1	5
2	4

    b.

Tableau 2

x	fréquence
6,3	2
6,4	1
6,5	6
6,6	5
6,7	13
6,8	4

    c.

Tableau 3

x	fréquence
1	15
2	5
3	3
4	1
5	2

3. Pour chacun des diagrammes à tiges et à feuilles suivants, trouvez :
    i. la moyenne;

ii. l'intervalle de classe modale.

a.

Tableau 4

Tige	Feuille
2	2 3 8
3	1 1 4 2
4	2 2 3 5 8 9 9
5	2 4 7 7 8
6	0 3 2
7	4

La tige 4 et la feuille 2 représente 42.

b.

Tableau 5

Tige	Feuille
$0^{(0)}$	2
$0^{(5)}$	5 6 8
$1^{(0)}$	0
$1^{(5)}$	5 5 6 6 7 8 8 9
$2^{(0)}$	0 0 0 1 1 2 3 3 3 4 4 4
$2^{(6)}$	6 6 7 8 8 9 9
$3^{(0)}$	0 4
$3^{(5)}$	5 6 7 7 8

La tige 3 et la feuille 5 représente 35.

4. Imaginez que les augmentations annuelles de la population durant une période de dix ans sont fournies dans le tableau ci-dessous :

Tableau 6. Accroissement de la population

Années	Augmentation par rapport à l'année précédente
1	53 377
2	52 170
3	67 000
4	90 332
5	72 681
6	65 226
7	76 777
8	83 657
9	77 753
10	82 892

a. Calculez l'augmentation annuelle moyenne de la population durant une période de dix ans.

b. Calculez l'augmentation annuelle médiane de la population durant une période de dix ans.

c. Croyez-vous que la différence entre ces deux mesures est significative? Expliquez votre réponse et quel résultat donne une meilleure indication du centre des données.

d. À quelles fins utiliserait-on des mesures comme celles mentionnées ci-dessus?

5. 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.

6. Imaginez que le nombre de chômeurs est fourni dans le tableau ci-dessous.

Tableau 7. Chômage

Groupe d'âge	Nombre de chômeurs
15 à 19	3 688
20 à 24	4 031
25 à 34	5 432
35 à 44	4 360
45 à 54	3 162
55 à 64	1 702

a. Copiez le tableau dans votre cahier et trouvez le point milieu de chaque intervalle. Calculez l'âge moyen d'une personne en chômage à l'aide du point milieu.

b. Quel est l'intervalle de classe modale?

c. Dans quel groupe d'âge la médiane se situe-t-elle?

d. Traitez brièvement de la comparaison de ces trois résultats.

e. Pourquoi pensez-vous que le nombre de chômeurs diminue après le groupe des 25 à 34 ans?

f. Comment les organismes d'aide sociale pourraient-ils utiliser ces chiffres?

7. Une enquête aléatoire de 100 hommes mariés a donné la distribution suivante d'heures qu'ils consacraient par semaine à un travail ménager non rémunéré :

Tableau 8. Heures consacrées par semaine à un travail ménager non rémunéré par 100 hommes mariés

Heure(s)	Nombre d'hommes
0 à < 5	1
5 à < 10	18
10 à < 15	24
15 à < 20	25
20 à < 25	18
25 à < 30	12
30 à < 35	1
35 à < 40	1

a. Copiez le tableau dans votre cahier et incluez dans le tableau des colonnes pour trouver l'extrémité (valeur la plus élevée) de chaque intervalle. Calculez les fréquences cumulées et les pourcentages cumulés, puis inscrivez-les dans votre tableau.

b. Dessinez l'ogive (ou la courbe de distribution) à l'aide de la fréquence cumulée de l'axe y.

c. À partir de la courbe, trouvez une valeur médiane approximative. Qu'est-ce que cette valeur indique?

d. Quel est l'intervalle de classe modale?

e. Calculez la moyenne. Qu'est-ce que cette valeur indique?

f. Décrivez brièvement la comparaison entre les valeurs moyenne, médiane et modale.

g. Comment détermineriez-vous si les femmes ont consacré plus d'heures par semaine que les hommes à un travail ménager non rémunéré?

8. Voici un tableau hypothétique de revenus annuels de personnes de 15 ans ou plus :

Tableau 9. Revenus annuels des personnes de 15 ans et plus

Revenu ($)	Nombre de personnes
0 à 2 079	114 195
2 080 à 4 159	44 817
4 160 à 6 239	45 862
6 240 à 8 319	139 611
8 320 à 10 399	114 192
10 400 à 15 599	148 276
15 600 à 20 799	123 638
20 800 à 25 999	121 623
26 000 à 31 199	103 402
31 200 à 36 399	73 463
36 400 à 41 599	59 126
41 600 à 51 999	68 747
52 000 à 77 999	56 710

a. Quel est l'intervalle de classe modale?

b. Copiez le tableau dans votre cahier et incluez dans le tableau des colonnes pour trouver l'extrémité supérieure de chaque intervalle. Calculez les fréquences cumulées et les pourcentages cumulés.

c. Dessinez l'ogive (ou la courbe de distribution).

d. À partir de la courbe, donnez une valeur approximative pour le revenu annuel médian hypothétique des personnes.

e. Calculez leur revenu annuel moyen hypothétique. (Tuyau : dans le tableau ci-dessus, l'intervalle 2 080 à 4 159 représente, en fait, 2 080 à < 4 160; le point milieu est donc 3 120.)

f. Comparez les valeurs moyenne, médiane et modale.

g. Quelle mesure donne l'indication la plus exacte du centre des données?

h. Quels types d'organisation utiliseraient ce genre d'information?

Activités en classe

1. Mesurez la taille, en l'arrondissant au centimètre près, de chaque élève de votre classe. Y a-t-il des valeurs aberrantes? Utilisez une méthode appropriée pour trouver la moyenne, la médiane et le mode. Comparez les trois mesures. Quelle valeur donne la meilleure mesure de tendance générale? Pourquoi? Quelles organisations ou entreprises trouveraient de telles statistiques utiles?

2. Calculez la population des élèves de votre année scolaire ou de votre école ces 10 dernières années. Y a-t-il des valeurs aberrantes? Utilisez une méthode appropriée pour trouver la moyenne, la médiane et le mode. Comparez les trois mesures. Quelle valeur donne la meilleure mesure de tendance centrale? Pourquoi? Comment de telles statistiques pourraient-elles être utilisées par votre école, votre commission ou votre conseil scolaire?

3. Trouvez les résultats finals du sport favori à votre école à partir des dossiers de l'établissement. Rassemblez les données sur les résultats, tant ceux des victoires que ceux des défaites, des 10 dernières années. (Si les données ne sont pas disponibles, utilisez de l'information relative à votre équipe sportive favorite.)
   - Quel a été le résultat final moyen des dix dernières années, en incluant aussi bien ceux des victoires que des défaites?
   - Quel a été le résultat final médian des dix dernières années, en incluant aussi bien ceux des victoires que des défaites?
   - L'un des résultats finals moyens est-il semblable au résultat final médian correspondant?
   - Compte tenu de ces valeurs, que peut-on dire au sujet des distributions?
   - Énumérez certains des problèmes que vous pourriez éprouver en essayant d'utiliser des statistiques pour comparer des équipes sportives de votre école ou d'autres équipes sportives du passé à des équipes d'aujourd'hui?
4. Pour des données ordinales, pouvez-vous citer des cas où le mode serait plus utile que la médiane ou que la moyenne? Discutez-en en classe.

Réponses

1.

a.

i. 0,1

ii. 0

iii. 0

b.

i. 2

ii. 2

iii. 2

c.

i. 2,8

ii. 2,5

iii. 3,9

d.

i. 154,3

ii. 154,3

iii. 152,3

2.

a.

i. 0

ii. 0

iii. 0

iv. La moyenne, la médiane et le mode sont égaux. Cette distribution est presque symétrique.

b.

i. 6,6

ii. 6,7

iii. 6,7

iv. La distribution est désaxée vers la gauche; la moyenne est donc inférieure à la médiane. Le mode et la médiane sont les mêmes. La médiane est une meilleure mesure de tendance centrale lorsque les distributions sont désaxées.

c.

i. 1,85

ii. 1

iii. 1

iv. La médiane et le mode sont les mêmes. La distribution est désaxée vers la droite; la moyenne est donc supérieure à la médiane. La médiane est une meilleure mesure de tendance centrale lorsque les distributions sont désaxées. Dans b) et c), la moyenne a été influencée par quelques valeurs faibles et d'autres élevées.

3.

a.

i. 48

ii. 40 à 49

b.

i. 23

ii. 20 à 24

4.

a. 72 186,5

b. 74 729

c. Les mesures sont assez rapprochées, étant donné la longueur de chaque observation. C'est la médiane qui donne probablement la meilleure indication du centre des données, puisqu'il y a une grande diversité de valeurs d'observation. La médiane ne serait influencée ni par les très grandes ni par les très petites valeurs.

d. Un gouvernement pourrait utiliser ces mesures lorsqu'il planifierait la construction d'écoles, d'hôpitaux et de routes. Il pourrait aussi s'en servir pour prévoir plus facilement ses recettes fiscales.

5.

a.

Tableau 1. Résultats d'un examen de mathématiques, pour lequel la note maximale qu'on pouvait obtenir était 10

Note (x)	Comptage	Fréquence (f)
0		
1	II	2
2	III	3
3	IIII	4
4	IIII	4
5	IIII	4
6	II	2
7	𝍸 𝍸	10
8	III	3
9	𝍸 I	6
10	II	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.

6.

a. 36,2

b. 25 à 34 (**Nota :** les longueurs des intervalles ne sont pas les mêmes. Si elles l'étaient, l'intervalle 15 à 24 serait l'intervalle de classe modale.)

c. 25 à 34

d. Les trois résultats se situent à l'intérieur du même intervalle, mais la distribution est désaxée (ou inclinée) vers la droite.

e. Les groupes d'âge plus jeunes, les 15 à 19 ans et les 20 à 24 ans, sont constitués de jeunes qui sont toujours aux études ou de diplômés qui sont incapables de se trouver un emploi. Les groupes d'âge après celui des 25 à 34 ans ont une plus faible proportion de personnes en chômage, puisqu'elles travaillent à plein temps et ne sont plus aux études.

f. Les organismes d'aide sociale pourraient utiliser ces chiffres pour planifier des programmes d'emploi qui s'adresseraient à des gens plus jeunes.

7.

a.

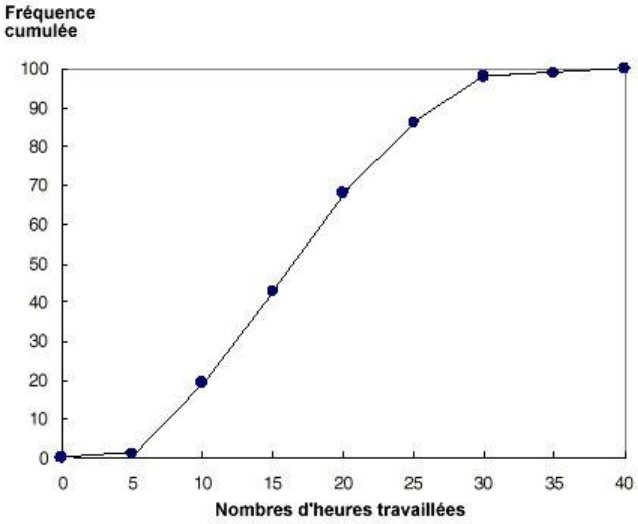
Tableau 2. Heures consacrées par semaine à un travail ménager non rémunéré

Heure(s)	Nombre d'hommes (x)	Extrémité	Fréquence cumulée	Pourcentage cumulé
0 à < 5	1	5	1	1
5 à < 10	18	10	19	19

10 à < 15	24	15	43	43
15 à < 20	25	20	68	68
20 à < 25	18	25	86	86
25 à < 30	12	30	98	98
30 à < 35	1	35	99	99
35 à < 40	1	40	100	100

b.

Figure 1. Heures consacrées par semaine à un travail ménager non rémunéré par 100 hommes mariés



c. La valeur approximative de la médiane est 17 heures, ce qui indique que le milieu de la distribution est 17 heures également.

d. L'intervalle de classe modale est 15 à < 20 heures.

e. La valeur moyenne est 16,8 heures, ce qui indique que le nombre moyen d'heures qu'un homme marié consacre à un travail ménager non rémunéré est 16,8 heures également.

f. La moyenne et la médiane sont très similaires et toutes les mesures se situent à l'intérieur de l'intervalle de classe modale. La distribution est presque symétrique.

g. On pourrait mener une enquête et en analyser les résultats de façon similaire. On pourrait ensuite comparer les résultats des deux enquêtes.

8.

a. L'intervalle de classe modale est 10 400 $ à 15 599 $. (**Nota :** les longueurs des intervalles ne sont pas les mêmes.)

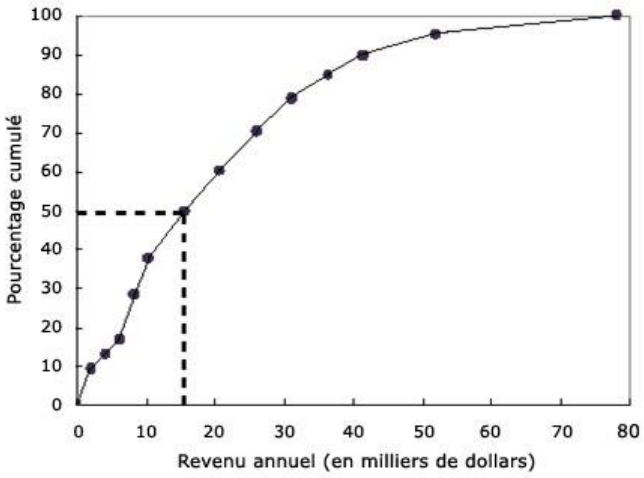
b.

Tableau 3. Revenus annuels des personnes de 15 ans et plus

Revenu ($)	Nombre de personnes	Extrémité	Fréquence cumulée	Pourcentage cumulé
		0	0	0,0
0 à < 2 080	114 195	2 080	114 195	9,4
2 080 à < 4 160	44 817	4 160	159 012	13,1
4 160 à < 6 240	45 862	6 240	204 874	16,9
6 240 à < 8 320	139 611	8 320	344 485	28,4
8 320 à < 10 400	114 192	10 400	458 677	37,8
10 400 à < 15 600	148 276	15 600	606 953	50,0
15 600 à < 20 800	123 638	20 800	730 591	60,2
20 800 à < 26 000	121 623	26 000	852 214	70,2
26 000 à < 31 200	103 402	31 200	955 616	78,7
31 200 à < 36 400	73 463	36 400	1 029 079	84,8
36 400 à < 41 600	59 126	41 600	1 088 205	89,7
41 600 à < 52 000	68 747	52 000	1 156 952	95,3
52 000 à < 78 000	56 710	78 000	1 213 662	100,0

c.

Figure 2. Revenus annuels hypothétiques de personnes de 15 ans et plus



d. Le revenu annuel médian de ces personnes est environ 15 500 $.

e. Le revenu annuel moyen est de 19 986 $.

f. Il est difficile de comparer le mode à la moyenne et à la médiane, en raison des différences qui existent entre les longueurs des intervalles. La moyenne est supérieure à la médiane, parce qu'elle est influencée par les revenus plus élevés. Cela signifie que la distribution est désaxée ou inclinée vers la droite.

g. La médiane donne l'indication la plus exacte du centre des données, parce qu'elle n'est pas influencée par des valeurs extrêmes.

h. Voici certaines des réponses possibles :

- des organismes d'aide sociale qui s'intéressent au nombre de travailleurs à faible revenu;
- des entreprises qui s'intéressent au nombre de travailleurs à revenu élevé;
- les gouvernements et d'autres fournisseurs de services qui s'intéressent aux données, ventilées suivant des caractéristiques comme l'âge, le sexe et le secteur géographique, afin d'offrir des services à des endroits appropriés

Mesures de dispersion

Les mesures de tendance centrale visent à identifier la valeur la plus représentative à l'intérieur d'un ensemble de données. Bien que la moyenne, la médiane et le mode donnent des perspectives différentes du centre d'un ensemble de données, une description de données n'est pas complète aussi longtemps qu'on ne connaît pas également la variabilité de sa distribution. En fait, la description numérique de base d'un ensemble de données exige des mesures tant du centre que de la dispersion. Les méthodes de mesure de dispersion incluent l' étendue , les quartiles , les écarts-types et la variance .

Étendue et quartiles

Étendue

L'étendue est très facile à calculer, parce qu'il s'agit simplement de la différence entre les valeurs observées les plus élevées et les plus faibles dans un ensemble de données. L'étendue, valeurs aberrantes comprises, est donc la dispersion réelle des données.

Étendue = différence entre les valeurs observées les plus élevées et les plus faibles

On ne tient pas compte de beaucoup de renseignements lorsqu'on calcule l'étendue, puisqu'on n'examine que les valeurs les plus élevées et les plus faibles.

La valeur de l'étendue d'un ensemble de données est grandement influencée par la présence d'une seule valeur inhabituellement élevée ou faible à l'intérieur de l'échantillon (une valeur aberrante).

On peut exprimer l'étendue sous la forme d'un intervalle comme 4 à 10, dans lequel 4 est la valeur la plus faible et 10, la valeur la plus élevée. On l'exprime souvent sous la forme de la longueur d'un intervalle. L'étendue de 4 à 10, par exemple, est de 6 chiffres. Nous utiliserons la dernière convention dans la présente section.

L'inconvénient d'utiliser l'étendue, c'est qu'elle ne mesure pas la dispersion de la majorité des valeurs d'un ensemble de données; elle ne mesure que la dispersion entre la valeur la plus élevée et la valeur la plus faible. Il faut donc d'autres mesures pour avoir une meilleure idée de la dispersion des données. L'étendue est un outil instructif qui sert de supplément à d'autres mesures comme l'écart-type ou l'écart semi-interquartile, mais on devrait rarement l'utiliser comme seule mesure de dispersion.

Quartiles

La médiane divise les données en deux ensembles égaux. (Pour plus de renseignements sur la médiane, consultez le chapitre sur les Mesuresdetendance centrale) :

- Le quartile inférieur est la valeur du milieu du premier ensemble, dans lequel 25 % des valeurs sont inférieures à Q_1 et 75 % lui sont supérieures. Le premier quartile prend la notation Q_1.
- Le quartile supérieur est la valeur du milieu du deuxième ensemble, dans lequel 75 % des valeurs sont inférieures à Q_3 et 25 % lui sont supérieures. Le troisième quartile prend donc la notation Q_3.

Il convient de noter que la médiane prend la notation Q_2, c'est-à-dire le deuxième quartile.

Exemple 1 – Quartiles supérieur et inférieur

- Données : 6, 47, 49, 15, 43, 41, 7, 39, 43, 41, 36
- Données ordonnées : 6, 7, 15, 36, 39, 41, 41, 43, 43, 47, 49
- Médiane : 41
- Quartile supérieur : 43
- Quartile inférieur : 15

Écart interquartile

L'écart interquartile est une autre étendue utilisée comme mesure de la dispersion. La différence entre les quartiles supérieur et inférieur (Q3 - Q1), qu'on appelle l'écart interquartile, indique aussi la dispersion d'un ensemble de données. L'écart interquartile couvre 50 % d'un ensemble de données et élimine l'influence des valeurs aberrantes, parce qu'on soustrait, en effet, le quartile le plus élevé et le quartile le plus faible.

Écart interquartile = différence entre le quartile supérieur (Q_3) et le quartile inférieur (Q_1)

Exemple 2 – Étendue et quartiles

Gabrielle a commencé à travailler dans une boutique d'informatique il y a un an. Son superviseur lui a demandé de tenir un dossier du nombre d'ordinateur(s) qu'elle a vendu(s) chaque mois.

L'ensemble de données qui suit indique le nombre d'ordinateur(s) qu'elle a vendu(s) mensuellement au cours des 12 derniers mois :

34, 47, 1, 15, 57, 24, 20, 11, 19, 50, 28, 37.

Utilisez les dossiers des ordinateurs vendus par Gabrielle pour trouver :

a. la médiane

b. l'étendue

c. les quartiles supérieur et inférieur

d. l'écart interquartile

Réponses

a. Les valeurs dans l'ordre ascendant sont :
1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57.

Quelle valeur prendre pour la médiane = (12e + 1er) ÷ 2
= $6,5^{e}$ valeur
= (6e + 7^{e} observations) ÷ 2
= (24 + 28) ÷ 2
= **26**

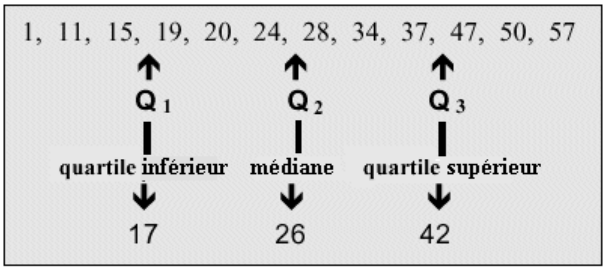
b. Étendue = différence entre la valeur la plus élevée et la valeur la plus faible.
= 57 – 1
= **56**

c. Quartile inférieur = valeur du milieu de la première moitié des données Q_1
= la médiane de 1, 11, 15, 19, 20, 24
= (3e + 4^{e} observations) ÷ 2
= (15 + 19) ÷ 2
= **17**

d. Quartile supérieur = valeur du milieu de la seconde moitié des données Q_3
= la médiane de 28, 34, 37, 47, 50, 57
= (3e + 4^{e} observations) ÷ 2
= (37 + 47) ÷ 2
= **42**

e. Écart interquartile = Q_3 – Q_1
= 42 – 17
= **25**

On peut résumer ces résultats comme suit :



Nota : Cet exemple renferme un nombre pair d'observations. La médiane, Q_2, se situe entre le centre de deux observations (24 et 28); le calcul de Q_1 inclut donc l'observation 24, puisqu'elle se situe au-dessous de la valeur de Q_2. De même, 28 est aussi inclus dans le calcul de Q3, puisqu'il se situe au-dessus de la valeur de Q_2.

Examinons un nombre impair d'observations comme 1, 2, 3, 4, 5, 6, 7. Ici, la valeur de Q_2 est 4. Comme la position de la médiane correspond exactement à la quatrième observation, cette valeur n'est pas incluse dans le calcul de Q_1 et de Q_3, étant donné que nous ne nous intéressons qu'aux données situées au-dessus et au-dessous de Q2. Dans l'exemple ci-dessus, Q_1 = 2 et Q_3 = 6.

Écart semi-interquartile

L'écart semi-interquartile est une autre mesure de dispersion. On le calcule sous la forme d'une moitié de la différence entre le 75e percentile (souvent appelé Q_3) et le 25e percentile (Q_1). La formule pour calculer l'écart semi-interquartile est la suivante :

(Q_3 – Q_1) ÷ 2.

Puisque la moitié des valeurs à l'intérieur d'une distribution se situe entre Q_3 et Q_1, l'écart semi-interquartile représente la moitié de la distance nécessaire pour englober 50 % des valeurs. Dans une distribution symétrique, un intervalle s'étirant d'un écart semi-interquartile au-dessous de la médiane à un écart semi-interquartile au-dessus de la médiane renfermera la moitié des valeurs. Toutefois, cela n'est pas vrai dans le cas d'une distributionasymétrique.

L'écart semi-interquartile n'est guère influencé par des valeurs plus élevées; c'est donc une bonne mesure de dispersion pour les distributions asymétriques. On utilise rarement des écarts semi-interquartiles pour des ensembles de données dont les distributions sont normales. Lorsqu'un ensemble de données comporte une distribution normale, on a plutôt recours à l'écart-type.

Variance et écart-type

Contrairement à l'étendue et aux quartiles, la variance permet de combiner toutes les valeurs à l'intérieur d'un ensemble de données afin d'obtenir la mesure de dispersion. La variance (symbolisée par $\mathbf{S^2}$) et l'écart-type (la racine carré de la variance, symbolisée par **S**) sont les mesures de dispersion les plus couramment utilisées.

Nous savons que la variance est une mesure du degré de dispersion d'un ensemble de données. On la calcule en prenant la moyenne de l'écart au carré de chaque nombre par rapport à la moyenne d'un ensemble de données. Pour les nombres 1, 2 et 3, par exemple, la moyenne est 2 et la variance, 0,667.

$[(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2] \div 3 = 0,667$

[somme de l'écart au carré] ÷ nombre d'observations = variance

Variance, (S^2) = moyenne de l'écart au carré de valeurs par rapport à la moyenne

Comme le calcul de la variance se fait à partir des carrés des écarts, les unités de mesure ne sont pas les mêmes que celles des observations originales. Par exemple, les longueurs mesurées en mètres (m) ont une variance mesurée en mètres carrés (m^2).

La racine carrée de la variance nous donne les unités utilisées dans l'échelle originale.

Écart-type (S) = Racine carrée de la variance

L'écart-type est la mesure de dispersion la plus couramment utilisée en statistique lorsqu'on emploie la moyenne pour calculer une tendance centrale. Il mesure donc la dispersion autour de la moyenne. En raison de ses liens étroits avec la moyenne, l'écart-type peut être grandement influencé si cette dernière donne une mauvaise mesure de tendance centrale.

L'écart-type est aussi influencé par les valeurs aberrantes; une seule de ces valeurs pourrait avoir une grande influence sur les résultats de l'écart-type. Il s'agit donc d'un bon indicateur de l'existence de valeurs aberrantes, ce qui en fait une mesure de dispersion très utile pour les distributions symétriques ne comptant aucune valeur aberrante.

L'écart-type est aussi utile quand on compare la dispersion de deux ensembles de données séparés qui ont approximativement la même moyenne. La dispersion des mesures autour de la moyenne est plus étroite dans le cas d'un ensemble de données dont l'écart-type est plus petit. Habituellement, 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.

Généralement, plus les valeurs sont largement distribuées, plus l'écart-type est élevé. Imaginez, par exemple, que nous devons séparer deux ensembles différents de résultats d'examens de 30 élèves; les notes du premier examen varient de 31 % à 98 % et celles du second, de 82 % à 93 %. Compte tenu de ces étendues, l'écart-type serait plus grand pour les résultats du premier examen.

Il n'est pas toujours facile d'évaluer l'importance que doit avoir l'écart-type pour que les données soient largement dispersées. L'importance de la valeur moyenne de l'ensemble des données dépend aussi de l'importance de l'écart-type. Lorsque vous mesurez quelque chose en millions, le fait d'avoir des mesures qui se rapprochent de la valeur moyenne n'a pas la même signification que si vous mesurez le poids de deux personnes. 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 quelle est l'importance de l'écart-type par rapport à la moyenne de l'ensemble de données.

Même s'il est moins sensible aux valeurs extrêmes que l'étendue, l'écart-type est quand même plus sensible que l'écartsemi-interquartile. Dans les cas où il y a des valeurs élevées aberrantes, on devrait aussi avoir recours à l'écart semi-quartile.

Propriétés de l'écart-type

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

- On n'utilise l'écart-type que pour mesurer la dispersion autour de la moyenne d'un ensemble de données.
- L'écart-type n'est jamais négatif.
- L'écart-type est sensible aux valeurs aberrantes. Une seule valeur aberrante peut accroître l'écart-type et, par le fait même, déformer le portrait de la dispersion.
- Dans le cas des données ayant approximativement la même moyenne, plus la dispersion est grande, plus l'écart-type est grand.
- L'écart-type est zéro si toutes les valeurs d'un ensemble de données sont les mêmes (parce que chaque valeur est égale à la moyenne).

Quand on analyse des données normalement distribuées, on peut utiliser l'écart-type parallèlement à la moyenne pour calculer des intervalles de données.

Si $\overline{x}$ = moyenne, **S** = écart-type et **x** = une valeur incluse dans l'ensemble de données, alors

- environ 68 % des données se situent à l'intérieur de l'intervalle : $\overline{x} - S < x < \overline{x} + S$.
- environ 95 % des données se situent à l'intérieur de l'intervalle : $\overline{x} - 2S < x < \overline{x} + 2S$.
- environ 99 % des données se situent à l'intérieur de l'intervalle : $\overline{x} - 3S < x < \overline{x} + 3S$.

Variables discrètes

On définit la variance d'une variable discrète composée de n observations comme suit :

$$S^2 = \frac{\sum (x - \overline{x})^2}{n}$$

L'écart-type d'une variable discrète composée de n observations est la racine carrée positive des variances et se définit comme suit :

$$S = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

Utilisez cette méthode étape par étape pour trouver l'écart-type d'une variable discrète.

1. Calculez la moyenne.
2. Soustrayez de chaque observation la moyenne.
3. Calculez le carré de chacune des autres observations.
4. Additionnez ces résultats au carré.
5. Divisez ce total par le nombre d'observations (la variance, S^2).
6. Utilisez la racine carrée positive (écart-type, **S**).

Exemple 1 – Écart-type

Une poule pond huit œufs. Voici les poids en grammes (g) des œufs :

60 g, 56 g, 6l g, 68 g, 51 g, 53 g, 69 g, 54 g.

a. Premièrement, calculez la moyenne :

$$\overline{x} = \frac{\sum x}{n}$$
$$= \frac{472}{8}$$
$$= 59$$

b. Maintenant, trouvez l'écart-type.

Tableau1. Poidsdesoeufs,engrammes

Poids (x)	$(x - \overline{x})$	$(x - \overline{x})^2$
60	1	1
56	-3	9
61	2	4
68	9	81
51	-8	64
53	-6	36
69	10	100
54	-5	25
472		**320**

À l'aide de l'information tirée du tableau ci-dessus, nous pouvons voir que :

$$\sum (x - \overline{x})^2 = 320$$

Pour calculer l'écart-type, on doit utiliser la formule qui suit :

$$S = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$
$$= \sqrt{\frac{320}{8}}$$
$$= 6{,}32 \text{ grammes}$$

Distribution de fréquences (variables discrètes)

Les formules pour la variance et l'écart-type changent légèrement si l'on groupe des observations à l'intérieur d'un tableau de fréquences. On multiplie les écarts au carré par la valeur de chaque fréquence, puis on calcule le total de ces résultats.

Dans une distributiondefréquences, on définit la variance pour une variable discrète comme suit :

$$S^2 = \frac{\sum (x - \bar{x})^2 f}{n} \qquad \text{où : } n = \sum f$$

L'écart-type d'une variable discrète se définit comme suit :

$$S = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n}}$$

Exemple 2 – Écart-type calculé à l'aide d'un tableau de fréquences

On a demandé à 30 fermiers combien de travailleurs agricoles ils embauchent durant des récoltes typiques. Voici leurs réponses :

4, 5, 6, 5, 3, 2, 8, 0, 4, 6, 7, 8, 4, 5, 7, 9, 8, 6, 7, 5, 5, 4, 2, 1, 9, 3, 3, 4, 6, 4

Tableau2. Nombredetravailleursagricolesembauchésdurantdesrécoltestypiquespar30fermiers

Travailleur (x)	Comptage	Fréquence (f)	(xf)	$(x - \bar{x})$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
0	I	1	0	-5	25	25
1	I	1	1	-4	16	16
2	II	2	4	-3	9	18
3	III	3	9	-2	4	12
4	IIII I	6	24	-1	1	6
5	IIII	5	25	0	0	0
6	IIII	4	24	1	1	4
7	III	3	21	2	4	12
8	III	3	24	3	9	27
9	II	2	18	4	16	32
		30	**150**			**152**

Pour calculer la moyenne :

$$\bar{x} = \frac{\sum x f}{\sum f}$$
$$= \frac{150}{30}$$
$$= 5$$

Pour calculer l'écart-type :

$$S = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n}}$$
$$= \sqrt{\frac{152}{30}}$$
$$= 2{,}25$$

Exemple 3 – Écart-type des variables groupées (continues ou discrètes)

On a demandé à un groupe de 220 élèves de 10e année combien d'heures ils ont regardé la télévision chaque semaine. Leurs réponses ont été consignées dans le tableau ci-dessous. À l'aide de cette information, calculez la moyenne et l'écart-type des heures pendant lesquelles les 220 élèves ont regardé la télévision.

Tableau 3. Nombre d'heures pendant lesquelles les 220élèvesontregardélatélévision

Heures	Nombre d'élèves
10 à 14	2
15 à 19	12
20 à 24	23
25 à 29	60
30 à 34	77
35 à 39	38

40 à 44	8

a. Premièrement, en utilisant le nombre d'élèves comme fréquence, trouvez le point milieu des intervalles de temps.

b. Calculez maintenant la moyenne à l'aide du point milieu (**x**) et de la fréquence (**f**).

Nota : Dans cet exemple, la variable continue utilisée a été arrondie au nombre entier le plus près. Le groupe **10 à 14** est en fait 9,5 à 14,499 (puisque 9,5 est arrondi à 10 et 14,499 à 14. La longueur de l'intervalle est 5, alors que le point milieu se situe à 12 (9,5 + 2,5 =12).

$$\bar{x} = \frac{\sum x f}{\sum f}$$
$$= \frac{6\,560}{220}$$
$$= 29{,}82$$

6 560 = (2 X 12 + 12 X 17 + 23 X 22 + 60 X 27 + 77 X 32 + 38 X 37 + 8 X 42)

Calculez ensuite les nombres pour les formules xf, $(x - \bar{x})$, $(x - \bar{x})^2$ et $(x - \bar{x})^2 f$.

Ajoutez-les au tableau de fréquences ci-dessous.

Tableau4.Nombred'heurespasséesdevantlatélévision

Heures	Point milieu (x)	Fréquence (f)	xf	$(x - \bar{x})$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
10 à 14	12	2	24	-17,82	317,6	635,2
15 à 19	17	12	204	-12,82	164,4	1 972,8
20 à 24	22	23	506	-7,82	61,2	1 407,6
25 à 29	27	60	1 620	-2,82	8,0	480,0
30 à 34	32	77	2 464	2,18	4,8	369,6
35 à 39	37	38	1 406	7,18	51,6	1 960,8
40 à 44	42	8	336	12,18	148,4	1 187,2
		220	**6 560**			**8 013,2**

Exemple 4 – Écart-type

Utilisez l'information fournie dans le tableau ci-dessus pour trouver l'écart-type.

$$S = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n}}$$
$$= \sqrt{\frac{8\,013{,}2}{220}}$$
$$= \sqrt{36{,}42}$$
$$= 6{,}03$$

Nota : Quand on groupe une variable par intervalle de classe, on suppose que toutes les observations à l'intérieur de chaque intervalle sont égales au point milieu de l'intervalle. Ainsi, on ne tient pas compte de la dispersion des observations à l'intérieur de chaque intervalle, ce qui fait que l'écart-type est toujours inférieur à la valeur réelle. On devrait donc le considérer comme une approximation.

Exemple 5 – Écart-type

En supposant que la distribution de fréquences est à peu près normale, calculez l'intervalle à l'intérieur duquel 95 % des observations incluses dans l'exemple précédent devraient se situer.

$\bar{x}$ = 29,82, **s** = 6,03

Calculez l'intervalle à l'aide de la formule suivante : $\bar{x}$ - **2s** < **x** < **2s** + $\bar{x}$

29,82 - (2 X 6,03) < **x** < 29,82 + (2 X 6,03)

29,82 - 12,06 < **x** < 29,82 + 12,06

17,76 < x < 41,88

Cela signifie une certitude d'environ 95 % qu'un élève passera entre 18 heures et 42 heures devant la télévision.

Résumé en cinq nombres

Un résumé en cinq nombres est surtout utile pour effectuer des analyses descriptives ou pour faire une analyse préliminaire d'un vaste ensemble de données. Un résumé se compose de cinq valeurs : les valeurs les plus extrêmes incluses dans l'ensemble de données (les valeurs maximale et minimale), les quartiles le moins et le plus élevés et la médiane. Ces valeurs sont présentées ensemble et ordonnées de la plus faible à la plus élevée : la valeur minimale, le quartile inférieur (Q_1), la valeur médiane (Q_2), le quartile supérieur (Q_3) et la valeur maximale.

On a sélectionné ces valeurs pour fournir un résumé d'un ensemble de données, parce que chaque valeur décrit une partie précise d'un tel ensemble : la médiane identifie le centre d'un ensemble de données; les quartiles supérieur et inférieur couvrent la moitié intermédiaire d'un ensemble de données; l'observation la plus faible et l'observation la plus élevée fournissent de l'information supplémentaire sur la dispersion réelle des données, ce qui fait que le résumé en cinq nombres est une mesure de dispersion importante.

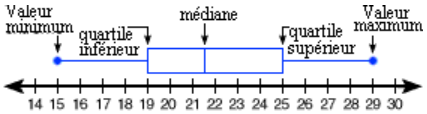
On peut représenter un résumé en cinq nombres à l'intérieur d'un diagramme appelé un diagramme de quartiles. Quand nous avons plus d'un ensemble de données à analyser, nous construisons pour chacun un résumé en cinq nombres et un diagramme de quartiles correspondant.

Pour illustrer notre propos, disons que le résumé en cinq nombres de l'exemple 2 donné dans la section sur l'étendue serait 1, 17, 26, 42, 57.

Construction de diagrammes de quartiles

Un diagramme de quartiles (parfois appelé un tracé en boîte ou un tracé en rectangle et moustaches) est un diagramme qui présente de l'information à partir d'un résumé en cinq nombres. Il ne montre pas de distribution aussi détaillée que le fait un histogramme ou un diagramme à tiges et à feuilles, mais il est surtout utile pour indiquer si une distribution est asymétrique et s'il y a des observations inhabituelles (des valeursaberrantes) dans l'ensemble de données. Les diagrammes de quartiles sont très utiles lorsqu'on a un grand nombre d'observations et qu'on doit comparer deux ensembles de données ou plus. (Voir la section sur les résumésencinqnombres pour plus de renseignements.)

Les diagrammes de quartiles sont idéals pour comparer des distributions, parce qu'ils font apparaître immédiatement le centre, la dispersion et l'étendue globale.

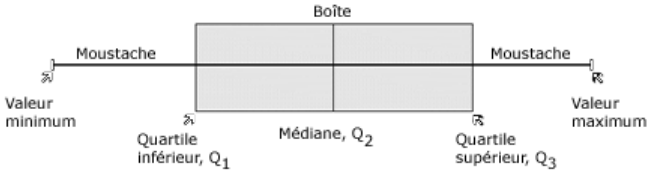


Le diagramme de quartiles est un moyen de résumer un ensemble de données mesuré à l'aide d'une échelle d'intervalles. On l'utilise souvent à l'intérieur d'une analyse explicative de données. Ce type de diagramme sert à montrer la forme de la distribution, sa valeur centrale et sa variabilité.

Dans un diagramme de quartiles :

- les extrémités du rectangle sont les quartiles supérieur et inférieur; le rectangle couvre donc l'écart interquartile;
- la médiane est indiquée par une ligne verticale située à l'intérieur du rectangle;
- les « moustaches » sont les deux lignes situées à l'extérieur du rectangle qui s'étendent de l'observation la plus élevée à l'observation la plus faible.

Figure 1. Diagramme de quartiles



Exemple 1 – Diagramme de quartiles

Comme Gabrielle, Samuel travaille dans une boutique d'informatique. Il a lui aussi enregistré le nombre d'ordinateurs qu'il a vendus chaque mois. Ces 12 derniers mois, il a vendu mensuellement le nombre d'ordinateurs suivants :

51, 17, 25, 39, 7, 49, 62, 41, 20, 6, 43, 13.

1. Donnez un résumé en cinq nombres du nombre d'ordinateurs que Samuel et Gabrielle ont vendus.
2. Créez deux diagrammes de quartiles, un pour le nombre d'ordinateurs vendus par Gabrielle et un pour celui des ordinateurs vendus par Samuel.
3. Décrivez brièvement les comparaisons entre les nombres d'ordinateurs qu'ils ont vendus.

Réponses

1. Premièrement, classez les données dans l'ordre ascendant. Trouvez ensuite la médiane.

   6, 7, 13, 17, 20, 25, 39, 41, 43, 49, 51, 62.
   Valeurs pour trouver la médiane = (12e + 1er) ÷ 2 = $6,5^{e}$ valeur
   = (6e + 7^{e} observations) ÷ 2
   = (25 + 39) ÷ 2
   = **32**

   Il y a six nombres au-dessous de la médiane : 6, 7, 13, 17, 20, 25.
   Q_1 = la médiane de ces six éléments
   = (6 + 1) ÷ 2= $3,5^{e}$ valeur
   = (3e + 4^{e} observations) ÷ 2
   = (13 + 17) ÷ 2
   = **15**

   Il y a six nombres au-dessus de la médiane : 39, 41, 43, 49, 51, 62.

Q_3 = la médiane de ces six éléments

= (6 + 1) ÷ 2= 3,5^{e} valeur

= (3e + 4^{e} observations) ÷ 2

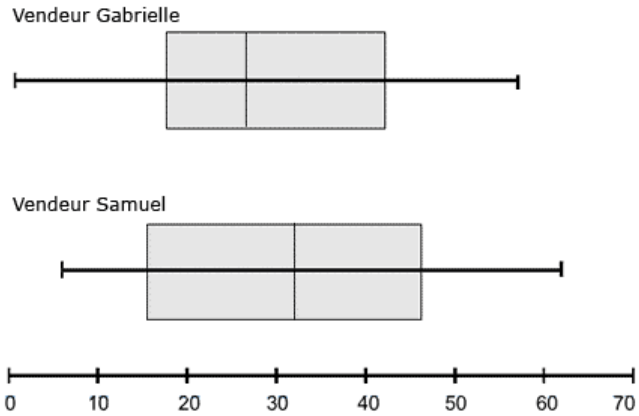
= **46**

Le résumé en cinq nombres pour le nombre d'ordinateurs vendus par Samuel est 6, 15, 32, 46, 62.

À l'aide des mêmes calculs, nous pouvons déterminer que le résumé en cinq nombres pour Gabrielle est 1, 17, 26, 42, 57.

2. Veuillez noter qu'on peut dessiner des diagrammes de quartiles verticalement ou horizontalement.

Figure 2. Tracés en rectangle et moustaches de Samuel et de Gabrielle



3. Le nombre le plus faible et le nombre le plus élevé d'ordinateurs vendus par Samuel sont tous les deux supérieurs aux nombres correspondants de ceux vendus par Gabrielle et le nombre médian d'ordinateurs vendus par Samuel est supérieur au nombre médian d'ordinateurs vendus par Gabrielle. De plus, l'écart interquartile de Samuel est plus grand que celui de Gabrielle.

   Ces résultats laissent entendre que Samuel vend constamment plus d'ordinateurs que Gabrielle.

Résumé

On peut décrire le centre et la dispersion d'une distribution de diverses façons. L'une de ces façons consiste à présenter l'information sous forme de résumé en cinq nombres. Le résumé, qui comporte une médiane comme valeur centrale, donne un aperçu des autres valeurs importantes à l'intérieur de la distribution. Il existe une autre mesure dans laquelle on utilise la moyenne et l'écart-type pour déchiffrer la dispersion des données. Toutefois, cette technique est plus utile dans le cas des distributions symétriques exemptes de valeurs aberrantes.

En dépit de cette restriction, les mesures comme la moyenne et l'écart–type sont utilisées plus couramment que le résumé en cinq nombres. Cela s'explique par le fait qu'on peut décrire approximativement un grand nombre de phénomènes naturels à partir d'une distribution normale. En ce qui concerne les distributions normales, la moyenne et l'écart-type sont considérés respectivement comme les meilleures mesures du centre et de la dispersion d'une distribution.

L'écart-type prend en compte toutes les valeurs d'une distribution, est extrêmement utile lorsqu'on l'utilise avec une distribution normale et, du point de vue mathématiques, est plus facile à utiliser. Toutefois, l'écart-type n'est pas une bonne mesure de dispersion lorsque les distributions sont très asymétriques. En pareil cas, on devrait aussi avoir recours à d'autres mesures comme l'écart semi-quartile.

On utilise rarement l'écart semi-interquartile comme mesure de dispersion, en partie parce qu'il est moins facile à utiliser que les autres mesures en usage. Malgré cela, cette statistique est utile puisqu'elle est moins influencée par des valeurs extrêmes que l'écart-type, est moins sujette aux fluctuations d'échantillonnage lorsque les distributions sont très asymétriques et se limite à deux valeurs, Q_1 et Q_2. L'écart semi-interquartile ne peut pas servir comme l'unique mesure de dispersion.

Exercices

1. Trouvez l'étendue pour les ensembles de données suivants.
    a. 6 – 8 – 11 – 15 – 24 – 38
    b. 11 – -6 – -2 – 16 – 9 – -8 – 17 – 19
    c. 6,4 – 3,8 – 5,9 – 4,7 – 5,3 – 7,1 – 3,2

2. Imaginez que le nombre de mariages enregistrés durant une période de dix ans est le suivant :

Tableau 1. Nombre de mariages enregistrés

Années	Nombre de mariages
1	40 650
2	40 812
3	41 300
4	41 450
5	39 594
6	40 734
7	39 993
8	38 814
9	37 828
10	35 716

À l'aide de l'information tirée du tableau ci-dessus, trouvez :
    a. l'étendue;
    b. la médiane;
    c. les quartiles supérieur et inférieur;
    d. l'écart interquartile;
    e. le résumé en cinq nombres.

3. Les températures énumérées ci-dessous sont les températures quotidiennes maximales (en degrés Celsius) enregistrées à Iqaluit du 2 au 16 juin :

2,8 – 7,3 – 9,6 – 8,9 – 11,4 – 6,7 – 5,8 – 5,5 – 6,7 – 6,2 – 9,0 – 8,2 – 7,6 – 8,5 – 6,7
    a. Trouvez l'étendue.
    b. Calculez l'écart interquartile.
    c. Donnez le résumé en cinq nombres.

4. Le tableau suivant fournit le nombre hypothétique de conflits de travail durant une période de dix ans.

Tableau 2. Nombre hypothétique de conflits de travail

Années	Nombre hypothétique de conflits de travail
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.

5. Voici le nombre de parties de basket-ball 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 diagramme à bandes verticales.

c. Calculez la moyenne, la médiane et le mode.

d. Calculez la variance et l'écart-type.

e. Calculez l'intervalle à l'intérieur duquel 95 % des observations devraient se situer.

f. Formulez un commentaire sur la dispersion des données.

Réponses

Note : Il existe différentes méthodes pour le calcul des quartiles. Par exemple, si on utilise les logiciels EXCEL ou SAS pour les calculs de ces exercices, les résultats pourraient être différents.

1.

a. 32

b. 27

c. 3,9

2.

a. 5 734

b. 40 321,5

c. Q_1 = 38 814 – Q_3 = 40 812

d. 1998

e. 35 716 – 38 814 – 40 321,5 – 40 812 – 41 450

3.

a. 8,6

b. 2,7

c. 2,8 – 6,2 – 7,3 – 8,9 – 11,4

4.

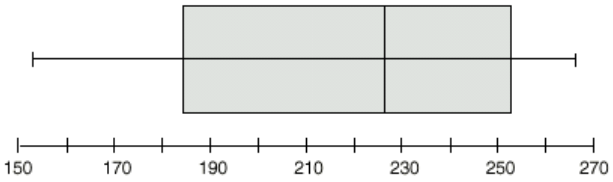
a. 113

b. 78

c. 153 – 182 – 226,5 – 260 – 266

d.

Figure 1. Nombre hypothétique de conflits de travail



5.

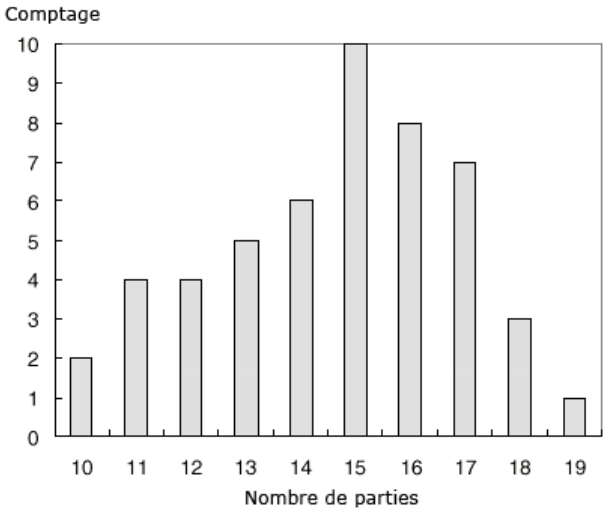
a.

Tableau 1. Nombre de parties de basket-ball auxquelles ont assisté 50 abonnés

Nombre de parties (x)	Comptage	Fréquence (f)
10	II	2
11	IIII	4
12	IIII	4
13	𝍸	5
14	𝍸 I	6
15	𝍸 𝍸	10
16	𝍸 III	8
17	𝍸 II	7
18	III	3
19	I	1
		50

b.

Figure 2. Nombre de parties de basket-ball auxquelles ont assisté 50 abonnés



c. moyenne = 14,62, médiane = 15, mode = 15

d. $\mathbf{S^2}$ = 4,96, S = 2,23

e. 10,16 < **x** < 19,08

f. L'écart-type est assez faible, ce qui indique que les données ne sont pas largement réparties autour de la moyenne. La moyenne et la médiane sont très rapprochées, ce qui indique que les données sont en gros symétriques.

Information : Utilisation dans la société

Vous connaissez sans doute l'expression *ère de l'information*. Elle sert à décrire dans quelle mesure l'information est une ressource essentielle dans la société moderne. Les gens se fondent sur l'information pour prendre des décisions et faire des recommandations dans de nombreux domaines, y compris la politique, l'économie, l'environnement et les divertissements.

Ce chapitre montre comment les personnes et les organisations utilisent l'information statistique. En l'absence de renseignements fiables, les gens peuvent prendre de mauvaises décisions qui ont parfois des conséquences graves.

Utilisation de l'information

Voici quelques principes de base à respecter lorsqu'on utilise des données statistiques. Veuillez garder présentes à l'esprit les lignes directrices ci-dessous :

- Allez à la recherche des faits. Déterminez quelle organisation ou quelles sources peuvent vous aider à trouver l'information dont vous avez besoin.
- Déterminez quel support répond le mieux à vos besoins. L'information est diffusée sous de nombreuses formes, y compris des publications imprimées et diffusées en direct, des CD-ROM et de plus en plus par l'entremise de sites Internet.
- Trouvez l'information. Si votre source est Statistique Canada, vous pouvez obtenir des données dans le site Web de Statistique Canada ou dans la plupart des bibliothèques communautaires. Si vous utilisez d'autres sources, assurez-vous que l'information est à jour et fiable. Les bibliothécaires de référence d'un bout à l'autre du pays pourront vous aider.
- Assurez-vous de comprendre les termes, les concepts et les définitions des données que vous utilisez, y compris la population observée et la méthodologie utilisée. Cela vous aidera à comparer des renseignements semblables provenant de différentes sources. Voici certains des termes statistiques courants : dollars courants et dollars constants, indices, arrondissement aléatoire et désaisonnalisation.

Voici quelques exemples des utilisations pratiques de l'information statistique :

- **Suivi de l'évolution de la part du marché :** Une société de fiducie et de prêts hypothécaires de Halifax souhaite évaluer sa part du marché des régimes d'épargne-retraite, des hypothèques et d'autres produits financiers clés. À l'aide des données des bilans, les chercheurs de la société peuvent comparer la croissance et la part de marché de cette dernière au total de la branche d'activité et aux chiffres des groupes financiers qui vendent les mêmes produits financiers. Ce faisant, ils déterminent les produits qui offrent les meilleures possibilités de nouvelles ventes. Ces faits d'importance capitale peuvent ensuite servir de fondement à une nouvelle stratégie de commercialisation.
- **Projection des frais de construction :** Un service public d'électricité prévoit construire une centrale électrique à Yellowknife. Les travaux prendront au moins dix ans. Les planificateurs ont besoin de connaître les coûts de construction futurs pour déterminer la viabilité du projet. La centrale électrique sera complexe et comprendra une section industrielle, une section commerciale et une section institutionnelle. Des consultations avec Statistique Canada permettront aux planificateurs de déterminer quels indices ils peuvent utiliser et (à supposer que l'économie demeure stable) de prévoir les changements des prix durant les dix prochaines années.
- **Suivi de l'évolution des prix de vente :** Une entreprise pharmaceutique de Winnipeg veut comparer l'évolution des prix de ses produits à ceux de l'ensemble de l'industrie. En utilisant une série d'indices pharmaceutiques, les analystes peuvent comparer l'évolution de leurs prix au fil du temps à ceux de l'industrie, de manière à déterminer la compétitivité de l'entreprise.
- **Prévision de l'évolution d'un marché :** Des fabricants de machines-outils de Sudbury ont besoin de connaître les dépenses prévues de l'industrie de l'automobile au titre des machines et du matériel. Les données sur les dépenses en immobilisations peuvent leur fournir les renseignements dont ils ont besoin pour planifier leurs investissements en nouvelles usines et en nouveau matériel.
- **Planification de la diversification :** Un grossiste de vêtements de Montréal qui a connu une réussite exceptionnelle souhaite mener des activités dans une autre industrie des services dans la province. Il examine d'abord les données sur les ventes au détail pour déterminer où les dépenses des consommateurs sont les plus élevées, et cela non seulement dans toute la province mais aussi à Montréal. L'indice montre aussi l'évolution des ventes au détail au fil des ans. À l'aide de cette information, le grossiste peut décider quel endroit offre les meilleures possibilités.
- **Une acquisition importante :** Les cadres d'une entreprise de l'industrie des ressources de la Colombie-Britannique examinent un certain nombre de sociétés en vue d'une éventuelle acquisition. Ils ont besoin de recueillir des renseignements sur la situation juridique, la structure et les filiales ainsi que les noms de tous les directeurs et agents de chaque société. Ils veulent aussi examiner la structure de plusieurs entreprises concurrentes pour déterminer dans quelle mesure elles sont intégrées dans l'industrie. Travaillant en étroite collaboration avec les préposés à l'application de la Loi sur les déclarations des personnes morales, ils accèdent à la Base de données sur les liens de parenté entre sociétés. Cette base de données contient une liste de tous les portefeuilles et de toutes les principales industries ainsi que les noms et adresses de tous les directeurs et agents.

Les données statistiques servent à de nombreuses autres fins. Pour vous renseigner sur ces utilisations, passez à la section sur les études de cas.

Étude de cas : L'appauvrissement de la couche d'ozone et le Protocole de Montréal

La couche d'ozone est une partie importante de l'atmosphère du globe et du système climatique mondial. Elle limite le rayonnement ultraviolet du soleil aux niveaux nécessaires à la vie sur terre. Une couche d'ozone appauvrie peut avoir des conséquences graves, y compris accroître l'incidence des coups de soleil, du cancer de la peau, des affections oculaires et d'autres maladies, en plus de réduire la croissance des plantes.

Certains composés chimiques fabriqués par l'homme sont les principaux responsables de l'appauvrissement de la couche d'ozone. Il s'agit de composés tels que les chlorofluorocarbures (CFC) et les halons, qui, auparavant, étaient couramment utilisés dans les réfrigérateurs, les climatiseurs et les substances ignifugeantes. De façon générale, une diminution de 1 % de l'ozone atmosphérique équivaut à une augmentation de 1 % à 2 % du rayonnement ultraviolet au niveau du sol.

État de la couche d'ozone

Depuis 1979, la couche d'ozone stratosphérique s'est amincie au-dessus de toute la surface du globe, de 4 % à 6 % par décennie aux latitudes moyennes et de 10 % à 12 % par décennie aux hautes latitudes méridionales. Les niveaux sont tombés à des planchers records à la suite de l'éruption du Mont Pinatubo, aux Philippines, en juin 1991. Toutefois, les répercussions de ce phénomène exceptionnel se sont atténuées et les niveaux ont retrouvé des valeurs plus proches de la tendance baissière à long terme.

Répercussions éventuelles de l'appauvrissement de la couche d'ozone

L'appauvrissement de l'ozone stratosphérique entraîne l'augmentation des rayons ultraviolets qui atteignent la surface de la terre. On sait qu'un niveau élevé de rayons UV ralentit la croissance des plantes et peut causer le cancer de la peau, des cataractes et des dommages au système immunitaire de l'homme et de certains animaux. Aux latitudes moyennes - par exemple celles de Toronto - et par temps clair, une réduction de 1 % de l'épaisseur de la couche d'ozone stratosphérique entraîne une augmentation d'environ 1,1 % à 1,4 % des rayons UV-B au niveau du sol. Cela varie selon la saison. Au Canada, environ 200 espèces de cultures et d'arbres sont sensibles, dans une certaine mesure, à l'accroissement des niveaux des rayons UV-B.

Mesures prises

Le Canada, l'un des premiers pays signataires du Protocole de Montréal de 1987, a joué un rôle de premier plan en étudiant les aspects scientifiques de l'appauvrissement de la couche d'ozone tout en prenant des mesures pour en éliminer les causes. La production de substances destructrices de l'ozone (SDO) au Canada est passée d'un sommet de 27,8 kilotonnes en 1987 à 1,0 kilotonne en 1996. À l'échelle mondiale, la production de CFC en 1995 était de 77 % inférieure au sommet atteint en 1988. Le Canada contribuait à moins de 1 % de la production globale en 1995.

Malgré les progrès accomplis, certaines préoccupations demeurent. En premier lieu, les chercheurs scientifiques ne peuvent encore confirmer avec certitude que, même si l'on atteint les objectifs d'élimination des SDO, la couche d'ozone retrouvera sa densité initiale. La concentration des SDO connues dans la stratosphère diminue, mais il peut y avoir d'autres substances qui contribuent à l'appauvrissement de la couche d'ozone. En deuxième lieu, les pays en voie de développement représentent maintenant une plus grande menace pour le rétablissement de la couche d'ozone, car leur production et leur utilisation des CFC ont augmenté au cours des dernières années. En troisième lieu, les efforts sont en perte de vitesse dans les pays développés, dont le Canada, car on a l'impression que le problème est maintenant réglé. Selon le rapport de 1997 du Vérificateur général du Canada, les stocks actuels de SDO qui se trouvent dans le matériel existant et les quantités entreposées pour utilisation ultérieure risquent d'être libérées dans l'atmosphère à moins qu'on ne mette en oeuvre des inspections et des mesures de protection plus rigoureuses et que ces mesures n'améliorent de plus de 10 % le rétablissement de la couche d'ozone.

Le problème de l'appauvrissement de la couche d'ozone est devenu évident dans les années 1980, quand des mesures scientifiques ont commencé à indiquer des baisses sensibles de la quantité d'ozone dans la stratosphère à l'échelle mondiale. On trouvera ci-dessous certains des résultats généraux.

- Aux latitudes moyennes, soit en Europe et en Amérique du Nord, on a enregistré des baisses annuelles moyennes de la quantité d'ozone allant de 2 % à 4 % au cours des années 1980.
- Pour l'Australie, les baisses enregistrées durant les années 1980 allaient de 0,5 % à 5 %.
- Pour ce qui est de l'Antarctique, le trou d'ozone au-dessus de l'hémisphère sud se forme chaque printemps et les pertes totales d'ozones enregistrées depuis 1985 atteignent entre 60 % et 70 %.

La gravité du problème a mené à une entente globale visant à réduire et à contrôler les émissions de SDO. En 1987, les représentants de 149 pays se sont réunis à Montréal et ont signé une entente visant à réduire l'utilisation de SDO. Voici les décisions qui ont été prises :

- Geler la consommation de CFC-11 aux niveaux de 1986 d'ici à 1989.
- Réduire la consommation de CFC-12 de 20 % d'ici le 1er juillet 1993.
- Mettre tout en oeuvre pour respecter les objectifs fixés à Montréal, exposés ci-dessous.

Tableau 1. Substances destructrices de l'ozone

	Potentiel de destruction de l'ozone[1]	Date d'élimination au Canada	Durée de vie dans l'atmosphère
Halons	3,0 à 10,0	1er janvier 1994	jusqu'à 65 ans
Tétrachlorure de carbone	1,1	1er janvier 1995	jusqu'à 42 ans
CFC	0,6 à 1,0	1er janvier 1996	de 50 à 1 700 ans
Méthylchloro-forme	0,1	1er janvier 1996	6 ans
Bromure de méthyle	0,6	1er janvier 2005[2]	jusqu'à 2 ans
HCFC[3]	0,001 à 0,52	1er janvier 2020	jusqu'à 19 ans

Notes :

1. Chaque substance qui appauvrit la couche d'ozone est mesurée selon un étalon normalisé appelé potentiel d'appauvrissement de la couche d'ozone (PDO). Pour plus de renseignements, consulter Environnement Canada, 1997, *Appauvrissement de l'ozone stratosphérique, Série nationale d'indicateurs environnementaux, Bulletin EDE No 97-2*, Ottawa.

2. Selon le règlement d'application de la *Loi de la protection sur l'environnement*, les SDO doivent être éliminées d'ici le 1er janvier 2010. Les pays en développement signataires du Protocole de Montréal élimineront le bromure de méthyle d'ici 2005.

3. La plupart des hydrochlorofluorocarbures ont été mis au point à titre de produits chimiques destinés à remplacer provisoirement les SDO les plus nocives, principalement les CFC.

Source : Vérificateur général du Canada, « *Protection de la couche d'ozone : le parcours inachevé* », Rapport du Vérificateur général du Canada à la Chambre des communes, en 1997, consulté le 14 juin 1999.

Étude de cas : Passages de la frontière entre les États-Unis et le Canada en voiture : Pourquoi ont-ils ralenti?

Dans cette étude de cas, vous comparerez le nombre d'automobiles qui sont entrées au Canada en provenance des États-Unis en passant par le pont international de Fort Erie en 2000 et en 2001. Les passages de la frontière étaient moins nombreux durant la première partie de 2000 qu'au cours de la première partie de 2001. Toutefois, les passages durant la période de septembre à décembre 2001 étaient moins nombreux qu'au cours de la même période de 2000.

Votre principal objectif dans cette étude de cas est de déterminer la raison de la différence importante dans le nombre de passages de la frontière, particulièrement entre septembre et décembre. Les événements survenus au début de 2000 différaient-ils de ceux survenus au début de 2001? Faisait-il mauvais à ce moment-là? Le prix de l'essence était-il élevé? Avant de continuer, faites un remue-méninges afin de trouver d'autres raisons qui pourraient expliquer la différence.

Même si cela peut sembler difficile, la première chose à faire par toute personne qui utilise des données est d'examiner les méthodes employées aux fins des calculs. Il faut procéder à cet examen dans tous les cas. Si on avait utilisé une nouvelle méthode pour compter les voitures l'année précédente, elle pourrait expliquer la différence entre les chiffres. Vous pouvez supposer ici que les mêmes méthodes ont été utilisées pour compter les voitures pour chacune des deux années, mais vous devriez quand même vérifier les méthodes de calcul employées dans tout travail à l'avenir.

Exercice 1

Faites cet exercice selon la marche à suivre ci-dessous pour accéder aux données dont vous aurez besoin aux fins de cette étude de cas.

1. Accédez au site Web E-STAT .
2. Acceptez la licence et à la page suivante, cliquez sur **Table des matières**.
3. À la section **La société**, cliquez sur le lien **Voyages et tourisme**.
4. Lorsque la nouvelle page s'affiche, choisissez **Voyages internationaux** dans la liste des sous-répertoires.
5. Cliquez sur le **Tableau 427-0002 (Nombre de véhicules voyageant entre le Canada et les États-Unis, données mensuelles)**.
6. Choisissez les options suivantes :
    - Géographie : choisissez **Fort Erie, Ontario**.
    - Caractéristiques du voyage : choisissez **Véhicules des États-Unis entrant au Canada – total**.
    - Durée du séjour : choisissez **Durée du séjour, total**.
    - Mode de transport : choisissez Automobiles.
    - Période : choisissez de **Janv. 2000** à **Déc. 2001**.
7. Choisissez l'option **Extraire séries chronologiques**.
8. Sous Spécification de la sortie, choisissez l'option **Texte, périodes = lignes**. N'apportez aucun autre changement. Cliquez sur **Extraire**
9. Entrez les résultats au tableau 1.

Tableau 1
Nombre de véhicules des États-Unis entrant par Fort Erie, Ontario, données mensuelles (toutes les durées de séjour) 2000 et 2001

	2000	2001
Janvier	126 847	142 810
Février		
Mars		
Avril		
Mai		
Juin		
Juillet		
Août		
Septembre		
Octobre		
Novembre		
Décembre		

Exercice 2

À l'aide des résultats indiqués au tableau 1, calculez la différence réelle en soustrayant la valeur pour 2000 de celle pour 2001. Calculez ensuite la variation en pourcentage du nombre total d'automobiles qui sont entrées par le poste frontalier de Fort Erie chaque mois entre 2000 et 2001, en exprimant le changement sous forme de pourcentage de la valeur pour 2000. Entrez les résultats au tableau 2.

Tableau 2
Variation mensuelle et variation en pourcentage du nombre de véhicules en provenance des États-Unis entrant par Fort Erie, Ontario (toutes les durées de séjour)

	2000	2001
Janvier	15 963	13,6
Février		
Mars		
Avril		
Mai		
Juin		
Juillet		
Août		
Septembre		
Octobre		
Novembre		
Décembre		

Exercice 3

Examinez les résultats consignés au tableau 2 et déterminez s'il y a eu une différence importante entre les quatre derniers mois de 2000 et la période correspondante en 2001. Pour chacune des années, indiquez parmi les quatres mois celui qui a connu la plus forte diminution du nombre de passages de la frontière en voiture?

Exercice 4

Le prix de l'essence pourrait être l'une des raisons des différences relevées et vérifiées dans l'exercice 3. Pour déterminer si les prix de l'essence étaient plus élevés pendant la période visée, visitez le site Web E-STAT de Statistique Canada. Naviguez vers les données sur les prix de l'essence. Ces prix ont-ils pu influer sur la baisse du nombre de voitures qui sont entrées au Canada en provenance des États-Unis au cours des quatre derniers mois de l'an 2000?

Exercice 5

À l'aide des données fournies dans le site Web de Statistique Canada, déterminez si le taux de change du dollar américain par rapport au dollar canadien est l'une des raisons qui expliquent le plus petit nombre de voitures qui sont entrées au Canada au cours des quatre derniers mois de 2000. Une marche à suivre pour accéder à ces données vous est proposée ci-dessous, mais n'hésitez pas à en trouver une qui vous est propre.

Dans la partie supérieure de la page du site Web de Statistique Canada, cliquez sur le lien Tableaux sommaires dans la barre de menus à gauche de la page d'acceuil.

Trouver le tableau Taux de change, taux d'intérêt, masse monétaire et prix des

actions. Cliquez sur le lien Banque du Canada au bas du tableau.

Lorsque la page s'affiche, choisissez Français, puis Autres taux de change.

Suivez le lien pour accéder à la page Moyennes mensuelles des taux de

change.

On peut consulter de nombreux taux de change. Pour savoir pourquoi on a enregistré une baisse dans le nombre de passages de la frontière, vous pourriez peut-être examiner les taux de change du dollar américain de 2000 à 2001.

Exercice 6

La baisse du nombre de passages à la frontière pourrait tenir au très mauvais temps qu'il faisait à la fin de 2001 par rapport au temps qu'il faisait au cours de la même période en 2000. Visitez le site d'Environnement Canada à l'adresse. Déterminez si le temps a joué un rôle dans la baisse du nombre de passages à la frontière, ou si d'autres facteurs imprévisibles ont pu avoir une incidence sur cette variation (p.ex., un événement historique).

Exercice 7

Dressez une liste d'autres raisons qui pourraient expliquer le nombre plus faible de passages de la frontière. Vous pouvez utiliser les données accessibles dans le site Web de Statistique Canada et peut-être d'autres sites pour déterminer si ces facteurs ont pu avoir une incidence.

Nota :

Vous pouvez aussi utiliser E-STAT pour confirmer vos calculs de la variation en pourcentage entre 2000 et 2001. Choisissez l'option **Spécification de la sortie**, et lorsque s'affichera le message « **la sortie contiendra** », cliquez sur la flèche descendante et choisissez « **la variation en pourcentage d'une année à l'autre** ».

Étude de cas : Les transactions bancaires électroniques – pas sûr qu'elles soient sûres?

Malgré l'augmentation du nombre de Canadiens qui surfent dans Internet ces dernières années, beaucoup ont encore de graves doutes quant à la sécurité des transactions bancaires électroniques (banque en ligne) et du commerce électronique (achat de biens et services en ligne). Souvent, les gens craignent d'entrer des renseignements personnels les concernant dans un réseau accessible au public comme le World Wide Web (WWW). Curieusement, toutefois, bon nombre de ces mêmes personnes n'hésitent pas à envoyer des messages électroniques contenant des renseignements personnels ou confidentiels.

Dans cette étude de cas, nous examinerons des données sur différentes sortes d'utilisations d'Internet et nous ferons des prévisions quant à l'évolution de ces chiffres dans un avenir rapproché. Nous examinerons certaines mesures de sécurité prises pour empêcher l'accès illicite aux renseignements personnels ainsi que les outils employés pour garantir la sûreté et la sécurité des renseignements confidentiels transmis sur Internet. Plus de gens feraient peut-être des transactions bancaires sur Internet s'ils connaissaient et comprenaient mieux les mesures de sécurité en place.

Exercice 1 – Pourcentage moyen d'utilisateurs d'Internet par ménage

1. D'après votre propre expérience et celle de votre famille et de vos amis, estimez le pourcentage de ménages canadiens qui utilisent Internet pour envoyer et recevoir des messages par courrier électronique, pour faire des transactions bancaires électroniques ou aux fins du commerce électronique. Consignez vos estimations.

2. Affichez les données sur les « Utilisation d'Internet par les individus, selon l'activité », en cliquant sur les Tableaux sommaires (disponible dans la marge de gauche de la page des ressources éducatives), puis en cherchant « Internet ». Assurez-vous de prendre les chiffres pour « Tous les Canadiens » dans la marge de gauche du tableau sommaire.Entrez les chiffres au tableau 1. Comparez les données diffusées pour la dernière année avec vos estimations. Vos estimations étaient-elles exactes?

Tableau 1. Pourcentage moyen de ménages qui utilisent Internet

Année	Courrier électronique	Transactions bancaires électroniques	Achat de biens et services

3. À l'aide des renseignements fournis au tableau 1, tracez des diagrammes linéaires pour chacune des trois activités proposées.

Exercice 2 – Croissance du courrier électronique à l'échelle mondiale

Le tableau 1 montre que le pourcentage de ménages qui utilisent Internet pour transmettre et recevoir du courrier électronique augmente. Dans quelques années, il atteindra près de 100 % pour les ménages qui utilisent régulièrement Internet à domicile et il ne serait pas étonnant si tous ceux qui ont accès à Internet utilisent le Web pour transmettre et recevoir des messages électroniques.

1. À l'aide de votre diagramme montrant le pourcentage de ménages canadiens qui utilisent Internet pour transmettre et recevoir du courrier électronique, faites une recherche pour trouver des données comparables sur d'autres pays et incluez-les dans votre diagramme.

   Voici une liste de sites Web qui pourraient vous aider dans vos recherches :

   - Bureaux de statistique nationaux
   - Organismes statistiques internationaux

2. Votre diagramme vous permettra de comparer la rapidité avec laquelle le Canada et d'autres pays s'habituent à utiliser Internet pour transmettre et recevoir le courrier électronique. Où pourriez-vous vérifier l'exactitude de votre diagramme?

Exercice 3 – Transactions bancaires électroniques et commerce électronique

Avec les données disponibles, il est difficile de préciser le pourcentage de ménages qui utilisent Internet aux fins de transactions bancaires électroniques ou de commerce électronique.

1. En prenant comme hypothèse la croissance linéaire des transactions bancaires électroniques et du commerce électronique, établissez l'équation de régression (droite de meilleur ajustement) pour chacun des deux ensembles de données. Au moyen de ces équations, déterminez quand le pourcentage d'utilisation d'Internet dans l'un et l'autre cas dépassera les 50 %.

2. Pour les besoins de cette question, prenez pour hypothèse que quiconque achète des produits en ligne fait aussi des transactions bancaires en ligne, mais que l'inverse ne se produit pas nécessairement. Remarquez l'écart assez constant ces dernière années entre les personnes qui font des transactions bancaires électroniques et celles qui achètent des produits en ligne.
    a. Les équations de régression calculées à la partie 1) de l'exercice 3 indiquent-t-elles que cet écart se maintiendra à environ 11 % au cours des prochaines années? À votre avis, les personnes qui font des transactions bancaires électroniques commenceront-elles à un moment donné à faire des achats en ligne?
    b. De façon générale, êtes-vous d'accord pour dire que les gens se sentent plus à l'aise en utilisant Internet pour faire des transactions bancaires électroniques que pour faire des achats en ligne?
    c. En pareil cas, cela vient-il confirmer l'hypothèse posée à la partie 2) de l'exercice 3 concernant les personnes qui font des achats en ligne et qui font aussi des transactions bancaires électroniques?

3. De nombreuses banques commencent à imposer des frais plus élevés aux clients qui préfèrent faire leurs transactions bancaires en personne. En même temps, ces institutions réduisent le nombre de leurs caissiers de banque.

4. Quelle incidence ces changements pourraient-ils avoir sur le pourcentage de personnes qui font des transactions bancaires ou des achats en ligne?

5. Quels autres facteurs ou événements pourraient avoir une incidence marquée sur le pourcentage de ces utilisateurs d'Internet?

Mesures de sécurité

Les préoccupations sur la plan de la sécurité sont l'une des principales raisons pour lesquelles certaines personnes n'utilisent pas Internet pour faire des transactions bancaires électroniques ou des achats en ligne. Curieusement, comme on l'a signalé, la transmission de renseignements confidentiels par courrier électronique ne semble pas inquiéter certaines de ces personnes. On semble croire que les messages transmis par courrier électronique sont plus sûrs et protégés que les transactions financières faites en ligne, point de vue qui semble peu logique. Cependant, cela soulève une question intéressante : comment les organisations empêchent-elles les criminels d'avoir accès aux renseignements confidentiels comme les cartes de crédit et les numéros de comptes bancaires?

Les entreprises, les banques et les organismes gouvernementaux utilisent de nombreux moyens pour protéger les renseignements personnels concernant leurs clients, y compris l'installation de matériel de détection des virus informatiques, de systèmes d'information de secours et de pare-feux (accès à Internet à sens unique). Toutes ces mesures de sécurité sont considérées comme des moyens efficaces d'empêcher l'accès non autorisé à un réseau informatique ainsi que les virus électroniques et les attaques pour refus de service. Le chiffrement des données est une autre mesure de sécurité répandue. Elle repose sur la cryptographie, l'étude de codes secrets et de chiffres, pour cacher l'information en créant des codes mathématiques.

Exercice 4 – Cryptographie

Faites une recherche sur les méthodes utilisées pour assurer la sécurité sur Internet. Par exemple, examinez comment la cryptographie utilise des nombres premiers pour assurer la sécurité et la protection des données.

Nous vous recommandons de lire un livre intitulé « In Code: A Mathematical Journey » (Workman Publishing, 2001). Il a été écrit par Sarah Flannery, une étudiante irlandaise âgée de 16 ans et son père, David Flannery. Dans ce livre, Sarah explique sa passion pour les mathématiques et raconte en détail ses recherches en cryptographie. Elle explique aussi les théories et les procédures du nouvel algorithme qu'elle a élaboré aux fins du chiffrement des messages électroniques. Pour plus de renseignements sur In Code et sur le travail de Sarah, consultez les sites Web suivants :

- « Mathematica and the Science of Secrecy »
- « The Factorization of RSA-129 »

Petit conseil! Si vous voulez en savoir davantage sur la cryptographie, faites une recherche sur Internet à l'aide du mot clé « RSA 129 ».

Exercice

1. Un soir, choisissez un canal de télévision canadien et regardez le téléjournal. Notez le sujet de tout reportage qui comprend des renseignements statistiques. Y a-t-il des décisions rapportées qui sont fondées sur des statistiques?

2. Examinez trois quotidiens nationaux portant la même date. Comptez le nombre d'articles à la première page de chacun. Combien de ces articles mentionnent des statistiques? Calculez le ratio entre les articles qui utilisent des statistiques et ceux qui n'en utilisent pas. Pour chaque quotidien, calculez le pourcentage d'articles en première page qui utilisent des statistiques. Présentez les résultats sous forme de tableau.

3. Imaginez que vous êtes un politicien qui veut faire baisser l'âge de voter de 18 à 16 ans. Quelles données statistiques pourriez-vous utiliser à l'appui de votre point de vue? Affirmeriez-vous que la décision de baisser l'âge de voter devrait être fondée uniquement sur les statistiques? Pourquoi, ou pourquoi pas?

4. Autres que les statistiques, quel genre de renseignements auriez-vous besoin pour prendre une décision sur l'une des questions ci-dessous? augmenter les impôts;
   - augmenter les impôts;
   - réduire l'encombrement des voies de circulation;
   - cesser de fumer;
   - acheter un ordinateur;
   - utiliser le réseau de transport public;
   - lancer une entreprise.

5. Quelle est votre ville préférée à l'étranger? Utilisez un élément d'information statistique pour appuyer votre réponse.

6. Y aurait-il des situations dans lesquelles la même information statistique pourrait servir à justifier des décisions contraires?

7. Rédigez un texte d'une page décrivant trois façons différentes d'utiliser l'information statistique dans la société. Donnez des exemples.

8. Examinez attentivement les chiffres de la population mondiale figurant au tableau 1 et répondez aux questions suivantes :
   - Quelles organisations ces chiffres pourraient-ils intéresser?
   - Quels enjeux ces chiffres permettraient-ils d'éclairer?
   - Dans le cadre de ces enjeux, quelles sortes de décisions prendriez-vous en tenant compte de l'information fournie dans le tableau ci-dessous? Discutez cette question en classe.

Tableau 1. Prévisions des Nations Unis sur la population mondiale

	1996	2050
	en millions de personnes	
Chine	1 232	1 517
Inde	945	1 533
Pakistan	140	357
Nigéria	115	338
Indonésie	200	318
Iran	70	170
É.-U.	269	347
Éthiopie	58	213
Brésil	161	243
Bangladesh	120	218
Kenya	28	66
Mexique	93	154
Fédération de Russie	148	114
Philippines	69	130
Ouganda	20	66

Source : Département de l'information économique et sociale et de l'analyse des politiques des Nations Unies. Population mondiale, 1996.

Problèmes liés à l'utilisation des renseignements

> Le jugement statistique sera un jour aussi nécessaire à l'exercice efficace des fonctions du citoyen que la capacité de lire et d ' écrire [Traduction]
> — *H.G. Wells*

La citation ci-dessus remonte au début du XXe siècle, et peu de gens seraient en désaccord avec sa pertinence aujourd'hui. Puisque les décisions qui touchent les citoyens sont souvent basées sur une forme quelconque de renseignements statistiques, il est important d' être au courant des problèmes possibles liés à la collecte, à l'interprétation et à la présentation de données statistiques. Il n'y a aucun doute qu'une mauvaise utilisation des statistiques , de même que d'autres formes de renseignements, contribue à fomenter le doute qu'a le public sur la validité de certains renseignements.

Ce chapitre donne un aperçu de certains problèmes auxquels vous pouvez être confronté si vous n'êtes pas vigilant au moment d'utiliser des statistiques.

L'utilisation efficiente et exacte de l' information est gênée par les défauts qui peuvent exister dans la structure statistique actuelle. Parfois, les données recherchées n'existent pas. Dans le même ordre d'idées, les données de différentes sources ne sont pas toujours comparables. De plus, il peut s'avérer nécessaire d'améliorer la qualité de certains renseignements. C'est pour ces raisons que l'utilisation de certaines données peut être limitée ou découragée.

Différents types d'utilisateurs se servent des renseignements statistiques de diverses façons. Avec un peu de prévoyance à l'étape de la planification, on peut rendre les données exactes, pertinentes et utiles pour une variété d'utilisateurs.

- Tout d'abord, il est important de déterminer l'utilisation finale des données.

  Par exemple, supposons que votre municipalité souhaite évaluer l'emplacement pour la construction d'une nouvelle autoroute. Les planificateurs de l'autoroute devraient prendre en considération ou recueillir plusieurs types de renseignements. Ils peuvent mener des enquêtes sur les incidences environnementales ou recueillir des données statistiques sur le coût moyen des terrains (afin de prévoir les coûts d'expropriation), les tendances démographiques, les prévisions des coûts de la construction et de la main-d'œuvre, de même que le nombre de résidents locaux. Bien que ces données appartiennent à différentes sous-catégories, les planificateurs doivent veiller à ce qu'elles soient exactes, appropriées et conformes à leur objet ultime : trouver le meilleur emplacement pour la nouvelle autoroute.

- Deuxièmement, l'actualité des données est également pertinente. Le type de données détermine la fréquence de leur collecte. Nos mesures financières nationales, notamment le produit intérieur brut, sont évaluées quatre fois par an, en trimestres (chaque année est divisée en quatre segments de trois mois). Par contre, le recensement — une évaluation complète de l'information démographique — est mené tous les cinq ans en raison de son coût, de sa portée et de sa vision à long terme. Comme la disponibilité des données a une incidence sur la période de référence d'une enquête, alors ce facteur doit également être pris en considération.

- Troisièmement, il est utile de classer les utilisateurs des données selon la nature de leurs activités. Cela aidera le statisticien et l'analyste à déterminer le type de renseignements nécessaires. Toutefois, le fait de savoir qui sont les utilisateurs n'aide pas nécessairement à connaître la manière dont ils se serviront des données.

- Enfin, il est important pour les utilisateurs de comprendre la méthodologie de l'enquête, y compris les concepts et définitions. Ainsi, ils comprendront mieux la portée de l'enquête et la façon appropriée d'utiliser les données.

Mauvaise interprétation des statistiques

La mauvaise interprétation des statistiques est un problème fréquent lié à l'utilisation de renseignements statistiques. Un certain nombre de facteurs peuvent être en cause.

Cette section explique comment les statistiques peuvent être mal interprétées en raison :

- d'une mauvaise compréhension des données,
- d'une utilisation de définitions non comparables,
- d'une mauvaise interprétation des renseignements en connaissance de cause.

Tous les jours, nous sommes bombardés de chiffres dans les médias. Bon nombre des faits et chiffres présentés dans les nouvelles, tels que les taux de divorce et de chômage, proviennent du *Quotidien*, le premier bulletin officiel de Statistique Canada qui comporte des données statistiques et des renseignements concernant les publications produites par Statistique Canada. Il présente l'analyse des données récemment diffusées et de l'information sur les sources pour des faits plus détaillés. Même si l'information provenant du bulletin Le Quotidien est objective, il est intéressant d'analyser la façon dont les journaux de différentes régions interprètent différemment les mêmes faits.

Mauvaise compréhension des données

Il arrive parfois que les données soient mal comprises par les médias. L'exemple 1 montre comment les tendances en matière de criminalité peuvent être exagérément simplifiées et mal interprétées.

Exemple 1 – Statistiques de la criminalité

Le Quotidien a publié des données sur les statistiques de la criminalité avec les sous-titres suivants dans le numéro qui suit :

Le Quotidien
Statistique Canada

Le Quotidien, le mardi 18 juillet 2000

Statistiques de la criminalité

Statistiques de la criminalité, 1999

- Diminution des crimes de violence dans toutes les catégories
- Forte diminution du nombre d'introductions par effraction
- Augmentation du nombre d'infractions reliées aux drogues, stabilité de celui des infractions de conduite avec facultés affaiblies
- Septième diminution consécutive de la criminalité chez les jeunes

Selon l'information contenue dans la publication tirée du Quotidien, les gros titres des journaux canadiens se lisent comme suit :

- **L'économie très rigoureuse empêche l'augmentation du taux d'homicide**
- **La diminution du taux d'homicide nuit aux plates-formes électorales**
- **Le taux d'homicide à son plus bas niveau en 32 ans**

Les gros titres de journaux susmentionnés se sont éloignés du contenu de l'article principal. En mettant l'accent sur un aspect des données, les journaux ont négligé les principales constatations, c'est-à-dire que le taux de criminalité national était en baisse pour la huitième année consécutive. De plus, les données ont révélé que le taux de crimes avec violence a diminué de 2,4 % en 1999, la septième diminution consécutive après 15 années d'augmentation. Toutes les principales catégories de crimes avec violence ont connu une baisse en 1999, y compris les homicides (-4,7 %), les tentatives de meurtre (-8,8 %), les voies de fait (-2,0 %), les agressions sexuelles (-7,3 %) et les vols qualifiés (-1,5 %).

Dans le bulletin, on mentionne également que le taux d'homicide national a généralement été en baisse depuis le milieu des années 1970. La tendance s'est maintenue en 1999; 536 homicides ont été déclarés par la police, 22 de moins que l'année précédente. Le taux d'homicide de 1999 — 1,76 homicides pour chaque tranche de 100 000 personnes — a été le moins élevé depuis 1967.

Bien qu'il soit possible de dire que la prospérité croissante puisse avoir contribué à un taux de criminalité moins élevé, les données fournies par Statistique Canada n'ont pas mesuré le lien entre les statistiques de l'économie et de la criminalité. Et si elles l'avaient mesuré, l'observation serait que le taux de criminalité a été en baisse depuis 1991, année où la récession au Canada a commencé.

On mentionne également que plusieurs facteurs contribuent aux changements du taux de criminalité. Toutefois, il semble que certains journalistes aient considéré cet énoncé comme un fait. Les connexions trouvées dans les gros titres des journaux sont des suppositions et n'ont pas été révélées dans les données objectives. Les journalistes ont mal interprété la publication des données, probablement parce qu'ils ont mal compris les causes et incidences sous-jacentes du crime.

Les représentants de Statistique Canada consacrent beaucoup de temps à l'examen de l'utilisation par les médias des données publiées chaque jour. Ces représentants répondent également aux questions des médias concernant les données et veillent à ce que les données soient adéquatement comprises. S'il y a malentendu, alors les représentants font de leur mieux pour le régler.

Utilisation de définitions non comparables

Il est important de comprendre les concepts et définitions statistiques qui sous-tendent les données que vous utilisez. Si vous examinez des questions de la population active, vous devriez vous familiariser avec les définitions des termes comme taux de chômage, taux d'emploi et taux de participation. Si vous examinez des données relatives aux questions environnementales, vous devrez prendre en considération la définition et les concepts associés à certains mots, notamment forêt, terre à bois, disparu, espèces en danger d'extinction et parc national.

Un grand avantage de l'information statistique est qu'elle peut être comparée, ce qui permet de révéler les tendances et caractéristiques. Par exemple, on peut comparer les conditions météorologiques de Vancouver avec celles d'Halifax, les résultats sportifs du passé avec ceux du présent ou le rendement scolaire des hommes avec celui des femmes.

Toutefois, la comparaison de statistiques peut entraîner des problèmes lorsque les définitions sous-jacentes, la classification ou les méthodes de collecte de données sont différentes. Cela est particulièrement vrai en ce qui concerne les statistiques provenant de sources différentes, et en particulier avec les statistiques de l'état civil. Veuillez vous référer au Tableau 1 ci-dessous.

Tableau 1
Personnes mariées selon le sexe, de 1998 à 2002

Nombre de personnes	Mariées				
	1998	1999	2000	2001	2002
Deux sexes	**14 630 173**	**14 711 793**	**14 806 694**	**14 913 766**	**15 018 130**
Hommes	7 299 132	7 337 226	7 381 266	7 431 522	7 476 537
Femmes	7 331 041	7 374 567	7 425 428	7 482 244	7 541 593

Source: Tableau CANSIM 051-0010.

Selon Statistique Canada, la définition de « personnes mariées » inclut les personnes qui sont légalement mariées vivant ensemble, les personnes qui sont légalement mariées et séparées, de même que les conjoints de fait (ou en « union libre »). Si vous compariez les chiffres du tableau avec ceux d'une enquête qui n'aurait pas inclu les conjoints de fait dans la catégorie « personnes mariées », les résultats seraient très différents, étant donné que les définitions ne sont pas les mêmes. Regardez les données du tableau ci-dessus. Pourquoi indiquent-elles qu'il y a plus de femmes mariées que d'hommes? Logiquement, les chiffres devraient être les mêmes. Il semble que chaque personne associe sa propre définition au terme « marié », ce qui fait que les chiffres sont légèrement différents.

Mauvaise interprétation des renseignements en connaissance de cause

Les renseignements factuels doivent être justes, objectifs et exacts. Malgré cela, il est important de reconnaître que les renseignements peuvent être mal interprétés en raison de points de vue tendancieux, de statistiques inexactes, et même en raison de l'ajout de données fictives.

Veuillez examiner l'énoncé ci-dessous :

> Les tacticiens politiques ne sont pas à la recherche de la pure vérité ou même de la simple exactitude. Ils sont à la recherche de munitions pour les guerres d'information. Les données, l'information et les connaissances n'ont pas besoin d'être vraies pour anéantir un adversaire.
> *—Alvin Toffler*

Toffler peut sembler exagérément cynique dans son point de vue, mais en réalité, les gens et les organisations manipulent réellement l'information pour leur propre usage. C'est pour cette raison que vous devriez toujours être critique à l'égard de l'information qui vous est communiquée. Assurez-vous de connaître la source de l'information et de savoir si cette source est crédible ou non. De plus, essayez de trouver les méthodologies qui ont servi à la collecte et au traitement des données.

Enfin, il est important de connaître le niveau d'exactitude des statistiques étant donné que les enquêtes sont sujettes à deux types d'erreurs : les erreurs d'échantillonnage et les erreurs non dues à l'échantillonnage. Les erreurs d'échantillonnage se produisent dans les enquêtes-échantillon parce qu'il y a seulement une partie de la population qui est étudiée, pas la population au complet. On retrouve les erreurs non dues à l'échantillonnage dans les enquêtes-échantillon comme dans les recensements.

Erreur d'échantillonnage

Lorsque vous menez une enquête-échantillon ,celle-ci sera sujette à ce que l'on appelle en statistique l'erreur d'échantillonnage.

L'erreur d'échantillonnage se produit lorsqu'on estime une caractéristique de la population en étudiant seulement une partie de la population au lieu de la population au complet. Il s'agit de la différence entre l'estimation calculée à partir d'une enquête-échantillon et la « vraie » valeur qui aurait été obtenue si un recensement auprès de la population entière avait été effectué dans les mêmes conditions. Dans un recensement, il n'y a pas d'erreur d'échantillonnage étant donné que les calculs sont basés sur la population au complet.

Caractéristiques

Erreur d'échantillonnage

- En général, l'erreur d'échantillonnage diminue au fur et à mesure que la taille de l'échantillon augmente (mais pas proportionnellement).
- L'erreur d'échantillonnage dépend de la taille de la population étudiée.
- L'erreur d'échantillonnage dépend de la variabilité de la caractéristique de la population qu'on étudie.
- On peut tenir compte de l'erreur d'échantillonnage et en réduire la portée grâce à un plan de sondage approprié.
- L'erreur d'échantillonnage peut être mesurée et contrôlée dans les enquêtes qui utilisent un échantillon probabiliste .

Taille de l'échantillon

En général, plus il y a de personnes qui participent à l'enquête (taille de l'échantillon), moins il y aura d'erreurs d'échantillonnage. Bon nombre de personnes sont étonnées de la petite taille d'enquêtes bien connues. Par exemple, les sondages qui tentent de prédire les tendances de vote sont menés à partir d'échantillons de 1 000 à 2 000 personnes, les échantillons d'environ 1 000 personnes étant les plus répandus. Les cotes d'écoute des programmes télévisés sont évaluées grâce à un sondage auprès d'environ 2 000 téléspectateurs. Ce petit échantillon représente les préférences à la télévision d'une population totale de 12 millions de ménages canadiens! Bien que beaucoup de gens pensent que de tels sondages sont fiables, certains statisticiens remettent en question leur exactitude, étant donné la petite taille de l'échantillon.

Si un des objectifs d'une enquête est d'étudier des sous-populations ou de mesurer des événements rares, alors un échantillon de plus grande taille sera nécessaire. Toutefois, il est important de noter qu'une augmentation de la taille de l'échantillon entraîne une augmentation des coûts.

Taille de la population

À l'exception des populations très petites où le lien est plus direct, la taille d'un échantillon n'augmente pas proportionnellement à la taille de la population. En fait, la taille de la population joue un rôle quasi inexistant lorsqu'il s'agit de grandes populations.

Variabilité de la caractéristique étudiée

En général, plus grande est la différence entre les unités de la population, plus grande sera la taille de l'échantillon nécessaire pour atteindre un certain niveau de fiabilité. Par exemple, si vous deviez mener une enquête sur les milieux de travail auprès d'une population dont les revenus varient entre 30 000 $ et 50 000 $, vous utiliseriez un échantillon de plus petite taille pour atteindre le même niveau de fiabilité que vous utiliseriez pour une population de taille égale pour laquelle les revenus varient entre 5 000 $ et 1 000 000 $.

Plan de sondage

Il est important d'élaborer un plan de sondage efficace, qui comporte un plan d'échantillonnage et une procédure d'estimation. La méthode ou plan d'échantillonnage peut grandement influer sur la taille de l'erreur d'échantillonnage. Bon nombre d'enquêtes comprennent des plans d'échantillonnage complexes. Ceux-ci entraînent souvent un plus grand nombre d'erreurs d'échantillonnage qu'un plan d'échantillonnage aléatoire simple . La procédure d'estimation a également une grande incidence sur l'erreur d'échantillonnage. On traite de ces concepts plus en détail dans le chapitre intitulé Méthodes d'échantillonnage .

Mesure des erreurs d'échantillonnage

Il existe des méthodes qui évaluent l'erreur d'échantillonnage pour les enquêtes portant sur un échantillon aléatoire. La variance d'échantillonnage est la mesure la plus couramment utilisée dans la quantification de l'erreur d'échantillonnage, et comme les autres méthodes, elle découle directement des méthodes d'échantillonnage et d'estimation utilisées dans l'enquête. (La variance d'échantillonnage est examinée plus en détail dans le chapitre intitulé Méthodes d'échantillonnage .)

Erreur non due à l'échantillonnage

Outre l'erreur d'échantillonnage associée au processus de sélection d'un échantillon, un large éventail d'erreurs peuvent être repérées dans une enquête. Ces erreurs sont habituellement appelées « erreurs non dues à l'échantillonnage ».

Les erreurs non dues à l'échantillonnage peuvent être définies comme étant des erreurs possibles pendant toutes les activités d'enquête, mis à part l'échantillonnage. Comparativement aux erreurs d'échantillonnage, les erreurs non dues à l'échantillonnage peuvent se retrouver aussi bien dans les enquêtes-échantillon que dans les recensements.

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

- Les **erreurs aléatoires** sont imprévisibles et découlent de l'estimation. Elles sont généralement éliminées si un échantillon suffisamment large est utilisé. Toutefois, lorsque ces erreurs ont lieu, elles entraînent souvent une variabilité accrue de la caractéristique d'intérêt (c.-à-d. plus grande est la différence entre les unités de la population, plus grande sera la taille de l'échantillon nécessaire pour atteindre un niveau particulier de fiabilité).
- Les **erreurs systématiques** ont tendance à s'accumuler dans tout l'échantillon. Par exemple, s'il y a une erreur dans la conception du questionnaire, cela pourrait causer des problèmes relativement aux réponses du répondant, qui pourraient à leur tour créer des erreurs de traitement, etc. Ces types d'erreurs peuvent souvent donner des résultats finals biaisés.

Les erreurs non dues à l'échantillonnage sont extrêmement difficiles sinon impossibles à mesurer. Étant donné que les erreurs aléatoires ont tendance à s'éliminer, les erreurs systématiques représentent la principale cause d'inquiétude. Comparativement au biais causé par la variance d'échantillonnage, celui causé par les erreurs systématiques ne diminue pas même si on augmente la taille de l'échantillon.

Caractéristiques

Les erreurs non dues à l'échantillonnage

- peuvent se produire dans tous les aspects du processus d'enquête, mis à part l'échantillonnage.
- se retrouvent dans les enquêtes-échantillon et les recensements.
- sont difficiles à mesurer.

Les erreurs non dues à l'échantillonnage peuvent survenir en raison de problèmes de couverture, de réponse, de non-réponse, de traitement des données, d'estimation et d'analyse. Chacun de ces types d'erreurs est expliqué ci-dessous.

Erreurs de couverture

Une erreur de couverture survient lorsqu'il y a une omission, une répétition ou un ajout erroné de certaines unités dans la population ou l'échantillon. Les omissions sont appelées « sous-dénombrement », tandis que les répétitions et les ajouts erronés sont appelés « surdénombrement ». Ces erreurs sont causées par des défauts dans la base de sondage : inexactitude, insuffisance ou répétition de données, obsolescence ou données inadéquates. Les erreurs de couverture peuvent également survenir lors de procédures sur place (p. ex., une enquête est menée, mais l'intervieweur oublie d'inclure plusieurs ménages ou personnes).

Erreurs de réponse

Les erreurs de réponse résultent des données qui ont été demandées, fournies, reçues ou enregistrées de façon erronée. Les erreurs de réponse peuvent survenir en raison d'inefficacités relativement au questionnaire, à l'intervieweur, au répondant ou au processus d'enquête.

- **Conception médiocre du questionnaire**
  Il est essentiel que les questions de l'enquête-échantillon ou du recensement soient soigneusement formulées de façon à éviter un biais. Si les questions induisent en erreur ou portent à confusion, alors les réponses peuvent être faussées.

  Pour plus de renseignements, veuillez vous référer à la section sur la conception des questionnaires.

- **Biais d'entrevue**
  Un intervieweur peut avoir de l'influence sur la façon dont un répondant répond aux questions d'une enquête. Cela peut se produire lorsque l'intervieweur est trop amical ou distant à l'égard du répondant, ou lorsqu'il pousse le répondant à répondre. Afin de prévenir cela, les intervieweurs doivent être informés pour pouvoir rester neutres tout au long de l'entrevue. Ils doivent également porter une grande attention à la façon dont ils posent chaque question. Si un intervieweur change la formulation d'une question, cela peut avoir une incidence sur la réponse du répondant.

- **Erreurs du répondant**
  Les répondants peuvent aussi donner de mauvaises réponses. Des souvenirs fautifs, la tendance à exagérer des événements ou à en minimiser l'importance, et la tendance à donner des réponses qui semblent plus « souhaitables sur le plan social » sont plusieurs raisons pour lesquelles un répondant peut donner une mauvaise réponse.

- **Problèmes relatifs au processus d'enquête**
  Des erreurs peuvent également survenir en raison d'un problème lié au processus d'enquête. Utiliser des réponses par procuration (prendre les réponses d'une personne autre que le répondant) ou manquer de contrôle sur les procédures d'enquête ne sont que quelques façons d'augmenter la possibilité d'erreurs de réponse.

Erreurs de non-réponse

Les erreurs de non-réponse se produisent lorsqu'on ne recueille pas suffisamment de réponses aux questions de l'enquête. Il y a deux types d'erreurs de non-réponse : les erreurs de non-réponse complète et partielle.

- **Erreurs de non-réponse complète**
  Ces erreurs se produisent lorsque les résultats omettent complètement les réponses de certaines unités de l'échantillon sélectionné. Les raisons derrière ce type d'erreur peuvent s'expliquer par le fait que le répondant n'est pas disponible ou est temporairement absent, qu'il est incapable de participer à l'enquête ou qu'il refuse, ou que le logement est vacant. Si un nombre important de personnes ne répondent pas à une enquête, alors les résultats peuvent être biaisés, étant donné que les caractéristiques des non-répondants peuvent différer des caractéristiques de ceux qui ont participé.

- **Erreurs de non-réponse partielle**
  Ce type d'erreur se produit lorsque l'information obtenue du répondant est incomplète. Certaines personnes peuvent avoir de la difficulté à comprendre certaines questions. Afin de réduire cette forme de biais, il faudrait porter une attention particulière au moment de concevoir et de mettre à l'essai les questionnaires. Des stratégies de vérification et d'imputation appropriées aideront également à minimiser ce biais.

Pour plus de renseignements sur la vérification et l'imputation, veuillez vous référer au chapitre intitulé Traitement des données.

Erreurs de traitement

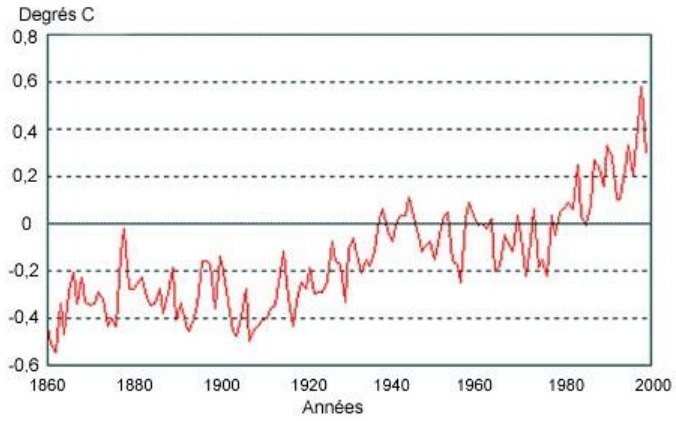
Les erreurs de traitement se produisent parfois pendant la préparation des fichiers de données finals. Par exemple, des erreurs peuvent survenir lorsque les données sont codées, saisies, corrigées ou imputées. Le biais du codeur est habituellement le résultat d'une formation médiocre ou d'instructions incomplètes, d'une variance du rendement du codeur (c.-à-d. fatigue, maladie), d'erreurs d'entrées de données ou d'un mauvais fonctionnement de la machine (certaines erreurs de traitement sont causées par des erreurs dans les programmes informatiques). On peut dire la même chose au sujet des erreurs saisies. Parfois, des erreurs sont identifiées à tort au cours de l'étape de vérification. Même lorsque des erreurs sont découvertes, elles peuvent être corrigées incorrectement en raison de procédures d'imputation médiocres.

Erreurs d'estimation

Statistique Canada et d'autres organismes de collecte de données consentent beaucoup d'effort à la conception et au suivi des enquêtes afin que ces dernières soient sans erreur dans la mesure du possible. Si on utilise une méthode d'estimation inappropriée, alors les résultats peuvent toujours être biaisés, malgré le fait que l'enquête était sans erreur avant l'estimation.

Voici un exemple d'une estimation possiblement inappropriée. Nous savons que le réchauffement de la terre est une question qui suscite beaucoup de débat. Pour s'assurer de bien mesurer ce phénomène, il faut être en mesure de trouver ce qui constitue une « moyenne de température planétaire » acceptable. La Figure 1 montre une représentation commune des données sur les changements climatiques. Elle indique une augmentation de la température moyenne du globe entre 0,3 ° et 0,6 °C sur approximativement 140 ans.

Figure 1 Changement du climat du globe, de 1860 à 2000



L'ensemble de données comprend des mesures qui ont été prises à diverses stations météorologiques partout dans le monde. Dans ce cas, la population est représentée par les mesures de la température à partir desquelles on tire un échantillon.

Certains scientifiques questionnent l'exactitude d'un diagramme comme la Figure 1, parce qu'ils croient que les estimations provenant de l'enquête-échantillon sont biaisées.

Ces scientifiques insistent sur le fait que les mesures de la température devraient refléter le ratio de la masse terrestre de la Terre par rapport à la masse d'eau. Par exemple, si la masse terrestre représente la moitié de la masse de l'eau (mers et océans), alors deux fois plus de mesures devraient provenir d'emplacements sur l'eau que d'emplacements sur terre. En fait, à la Figure 1, peu de mesures ont été prises à partir d'emplacements à la surface de l'eau, tandis que la grande majorité des mesures ont été prises à partir de stations météorologiques sur terre.

Pourquoi cela pourrait-il fausser les estimations de l'enquête-échantillon?

Les températures sur la terre ont tendance à être naturellement plus élevées que celles à la surface de l'eau, à cause du phénomène connu sous le nom d'« effet de l'îlot thermique urbain ». Si l'échantillon est trop lourdement pondéré en faveur des températures basées au sol, et que les estimations ne prennent pas cela en compte (ce que certains scientifiques déclarent), alors les résultats peuvent ne pas refléter réellement une moyenne mondiale.

Pour plus de renseignements, consultez la section sur l'estimation.

Erreurs d’analyse

Les erreurs d’analyse comprennent toutes celles qui se produisent lorsqu’on utilise les mauvais outils d’analyse ou lorsqu’on fourni les résultats préliminaires au lieu des résultats finals. Les erreurs qui se produisent pendant la publication des résultats sont également considérées comme des erreurs d’analyse.

Résumé

Lorsqu'on vous présente des renseignements statistiques, il est utile d'avoir une liste de questions prêtes pour vous aider à estimer la fiabilité des statistiques. Il ne s'agit pas de suggérer qu'il y a des problèmes avec les données, mais de vous aider à avoir plus d'assurance dans vos interrogations sur la fiabilité des données.

Les questions suivantes sont simplement des questions que vous devriez poser lorsqu'on vous présente des renseignements statistiques. Rappelez-vous que si la source ne peut vous donner de réponses ou d'explications, alors vous devriez vous demander si les données sont vraiment valables.

- Quel est l'auteur (source) des renseignements?
- S'agit-il de la source principale (c.-à-d. l'organisation qui a recueilli les données) ou d'une source secondaire (une organisation ou un analyste de l'extérieur)?
- Est-ce que la principale source des renseignements a une raison de mal interpréter les renseignements?
- Si les renseignements proviennent d'une source secondaire, est-il possible que les données aient été modifiées pour une raison quelconque?
- Est-il nécessaire de trouver la méthode de collecte des données, la technique d'échantillonnage ou le taux de réponse à l'enquête?
- Si les renseignements sont tirés d'une enquête-échantillon, croyez-vous que la taille de l'échantillon était adéquate? Quel est le niveau d'erreur d'échantillonnage?
- Les questions de l'enquête étaient-elles faciles à comprendre?
- Comprenez-vous les définitions des variables ou des sujets discutés dans l'enquête ou le recensement?
- Les définitions sont-elles uniformes?

Le lien *Trouver les statistiques et les utiliser* vous renseignera sur la façon de repérer de l'information fiable et d'en tirer parti. Ce lien permet de savoir comment trouver les données voulues, par où commencer, comment interpréter les tableaux statistiques, les symboles et définitions, comment mettre les données en contexte, utiliser les techniques de base, et enfin, tirer des conclusions.

Exercices

1. Déterminez lequel des énoncés suivants est fiable. À l'aide des connaissances que vous avez acquises dans ce chapitre, expliquez les problèmes possibles liés à chaque énoncé.
    a. Le revenu annuel moyen d'un Canadien est de 29 100 $, selon une enquête menée à Oshawa.
    b. Une grande majorité de personnes des régions rurales appuient les subventions pour les exploitations agricoles en difficulté. Il s'agit du résultat d'un sondage téléphonique à ligne ouverte mené par une station de télévision régionale.
    c. Les statistiques révèlent que 30 % des décrocheurs de notre pays se situent au dessous de la moyenne relativement aux capacités de lecture et d'écriture.
    d. Au Canada, le taux de chômage des jeunes représente plus de 12 %; donc plus de 12 % des Canadiens entre 15 et 24 ans sont sans emploi.
    e. Récemment, un important groupe environnementaliste a déclaré que seulement 3 % de la couverture terrestre du Canada était constituée de forêt, tandis qu'une importante entreprise a déclaré qu'il s'agissait plutôt de 6 %.
2. Quels sont les deux types d'erreurs non dues à l'échantillonnage?
3. Par quoi l'erreur d'échantillonnage est-elle influencée?
4. Dressez la liste des sources possibles d'erreurs systématiques (ou de biais).

Activité de groupe

1. Menez un débat en classe sur l'énoncé suivant :
    - Ce n'est pas assez d'exprimer les données statistiques seulement au moyen de tableaux et de diagrammes — on doit appuyer les données au moyen d'un texte.

Réponses

1.

a. Non, cet énoncé n'est pas fiable. Cette statistique est le résultat d'une estimation inappropriée en raison de son échantillon non représentatif. Le fait de choisir Oshawa à titre de population représentative ne tient pas compte des variations régionales telles que le type d'industrie (p. ex. relative aux ressources, tourisme, fabrication, haute technologie), ou d'emploi (p. ex. saisonnier, permanent). Afin d'avoir un échantillon représentatif, le statisticien doit recueillir des renseignements d'un bon nombre de secteurs différents à l'échelle du pays, y compris les régions rurales, semi-urbaines et urbaines.

b. Non, cet énoncé n'est pas fiable. Les participants à un sondage téléphonique à ligne ouverte d'une station de télévision se sélectionnent de façon volontaire, ce qui peut mener à un échantillon non représentatif.

c. Cet énoncé peut être fiable ou pas. Sa fiabilité est remise en cause parce que nous ne savons pas d'où proviennent les données statistiques sur les décrocheurs scolaires.

d. Encore une fois, cet énoncé peut être fiable ou pas. Nous ne connaissons pas l'origine de ces renseignements. Toutefois, si l'enquête a été menée de façon appropriée à l'échelle nationale, et si la définition de « jeunes » comprend les personnes âgées entre 15 et 24 ans, alors il n'y a pas de raison de dire que cette statistique est incorrecte.

e. Non, cet énoncé n'est pas fiable. Si les statistiques sont conflictuelles, cela doit signifier qu'on parle de différentes définitions de « couverture terrestre constituée de forêts » ou bien qu'une des sources est fautive ou bien que les résultats ont été mal interprétés par une des parties.

2. Il y a deux types d'erreurs non dues à l'échantillonnage : les erreurs aléatoires et les erreurs systématiques (ou biais).
3. L'erreur d'échantillonnage est influencée par la taille de la population et de l'échantillon, la variabilité de la caractéristique d'intérêt de la population, le plan d'échantillon et la méthode d'estimation.
4. Les sources suivantes peuvent donner lieu à des erreurs systématiques :
   - erreurs de couverture
   - erreurs de réponse
   - erreurs de non-réponse
   - erreurs de traitement
   - erreurs d'estimation
   - erreurs d'analyse.

Confidentialité, protection des renseignements personnels et sécurité

À mesure que l'information prend une importance capitale dans notre société, de nouveaux problèmes se posent sur le plan des droits individuels et de la protection des renseignements personnels. Les gens veulent toutes sortes de renseignements, comme les derniers chiffres sur les emplois obtenus par les jeunes qui viennent de terminer leur programme d'études et les soldes exacts des comptes bancaires fournis par les guichets automatiques. Parallèlement, de nombreuses personnes s'inquiètent de ce que trop de renseignements à leur sujet soient stockés dans des bases de données informatiques et que des personnes ou des organisations qu'ils ne connaissent pas y aient accès.

Les écrivains et les cinéastes ont brossé des tableaux terrifiants de sociétés futuristes dans lesquelles des ordinateurs tout-puissants contrôlent les pensées et les actes des êtres humains. Ces craintes sont souvent exagérées dans l'intérêt d'un bon récit. Toutefois, la société doit se pencher sur certaines questions importantes de façon à pouvoir garantir la protection des droits de l'individu à l'ère de l'information. Le présent chapitre porte sur plusieurs de ces questions.

La fourniture de renseignements

Les gens sont constamment appelés à fournir des renseignements sur eux-mêmes. Pour louer une vidéo ou sortir un livre de la bibliothèque, il faut remplir une formule de renseignements personnels. Les contribuables et les utilisateurs de services gouvernementaux doivent fournir des détails personnels.

De même, les banques et les gros magasins de détail ne donnent une carte de crédit à un client que s'ils ont des renseignements sur son revenu, sa profession et sa situation de famille. Les services de santé recueillent et emmagasinent des données sur chaque patient traité.

Ce ne sont là que quelques exemples de situations où les gens doivent fournir des renseignements personnels, dont une bonne partie sont entrés dans des bases de données informatiques. Le fait que les autorités et d'autres organisations pourraient accéder et se relier à ces bases de données est une source d'inquiétude. On pourrait ainsi établir des profils de données ou des fiches descriptives de particuliers pouvant être utilisés d'une manière préjudiciable.

Pour protéger la vie privée, on ne doit utiliser les renseignements personnels qu'aux fins auxquelles leur collecte a été autorisée.

L'atteinte à la confidentialité des renseignements des particuliers et des sociétés par la pénétration de bases de données est également une source d'inquiétude. Ce piratage informatique est habituellement perpétré par des utilisateurs qui obtiennent un accès non autorisé à d'importantes bases de données et qui peuvent même modifier les fichiers. De pareilles activités sont tout à fait illégales.

Si les inquiétudes au sujet de la protection des renseignements personnels sont légitimes, la fourniture d'information est utile à bien des égards. La société moderne a besoin d'information pour fonctionner efficacement. D'autre part, les gens qui fournissent des renseignements doivent avoir la certitude que la sécurité et le caractère confidentiel des données les concernant sont protégés.

Protection des renseignements personnels et sécurité

Les gens ont besoin de plus en plus d'information et de compétences en traitement de celle-ci pour prendre des décisions. En cette ère de l'information, la protection des renseignements personnels demeure une question d'importance capitale.

Au Canada, les organismes gouvernementaux et communautaires sont toujours à la recherche de moyens pour mieux protéger le caractère confidentiel des renseignements statistiques. Par exemple :

- les détenteurs de données sont encouragés à appliquer des procédures de sécurité rigoureuses afin que seules les personnes autorisées puissent accéder aux bases de données;
- des liens de télécommunication beaucoup plus sûrs sont maintenant disponibles pour la transmission de données entre différents secteurs d'une entreprise ou d'un organisme gouvernemental donné;
- des procédures plus strictes sont en place pour empêcher la prolifération des virus informatiques.

Les valeurs de Statistique Canada

> *« Les notions de fiabilité, d'objectivité et de confidentialité vont de pair et sont primordiales dans le milieu de la statistique et le fonctionnement d'un organisme statistique. »*
> *— M. Martin B. Wilk, statisticien en chef, 1980 à 1985*

À Statistique Canada, la protection et la sécurité des renseignements fournis demeure une préoccupation de la plus haute importance. Pour s'assurer que cette préoccupation est prise en considération en tout temps, nous avons élaboré une série de valeurs dans le but d'atteindre notre objectif, que voici :

- demeurer objectif; protéger
- la confidentialité; se
- concentrer sur l'analyse;
- alléger le fardeau de réponse;

- maintenir des normes de professionnalisme et de fiabilité.

La protection des renseignements personnels et la sécurité à Statistique Canada

En sa qualité de bureau fédéral de la statistique, Statistique Canada prend des mesures vigoureuses pour assurer la confidentialité et la sécurité des données fournies par les particuliers, les entreprises et les organisations.

Tous les employés de Statistique Canada prêtent un serment de discrétion et s'exposent à des peines sévères pour toute violation du secret professionnel. Les employés qui ne respectent pas leur serment sont passibles d'amendes et/ou d'une peine d'emprisonnement de six mois.

Les renseignements recueillis en vertu de la *Loi sur la statistique* ne peuvent être divulgués en application de la *Loi sur l'accès à l'information* ni de toute autre loi. L'Agence canadienne des douanes et du revenu, la Gendarmerie royale du Canada et les tribunaux n'ont pas accès aux réponses fournies dans le cadre des enquêtes.

Contrôle strict de l'accès

Tous les employés de Statistique Canada ont la responsabilité d'assurer la sécurité des données confidentielles. Seul le personnel qui a besoin de consulter des dossiers confidentiels dans l'exercice de ses fonctions est autorisé à y avoir accès. Il existe tout un réseau de systèmes et de procédures de sécurité matérielle visant à protéger les données confidentielles contre tout accès non autorisé.

La protection électronique des données confidentielles conservées dans les bases de données informatiques est un autre aspect important du système de sécurité de Statistique Canada. Le Bureau a établi deux réseaux distincts, l'un interne, l'autre externe. La conservation, le transfert et le traitement des données confidentielles se limitent au seul réseau interne, ce qui empêche tout accès de l'extérieur.

Protection contre la divulgation

Statistique Canada publie les données sous forme de résumés, de tableaux et de diagrammes statistiques. Aucune donnée ne peut être publiée si elle risque d'identifier un particulier, une entreprise ou une organisation, à moins que ce particulier, cette entreprise ou cette organisation n'en ait connaissance ou ne donne son consentement. Le Bureau filtre soigneusement tous les résultats définitifs avant de les diffuser. Par exemple, si seulement deux compagnies dans une province fournissent un produit donné, nous ne diffuserons pas les données sur les ventes de ce produit pour cette province sans le consentement écrit des deux compagnies. Cette restriction s'applique également si trois compagnies fournissent le produit et qu'une ou deux d'entre elles dominent l'industrie dans cette province. Si le chiffre pour une province n'est pas publié, le chiffre pour au moins une autre province est supprimé afin d'éviter la possibilité que quelqu'un ne détermine le chiffre manquant à partir du total national.

En plus de la technique de suppression, on utilise d'autres méthodes pour éliminer toute possibilité de divulgation. En voici quelques-unes : l'arrondissement, le regroupement, la micro-agrégation et la permutation des données.

Renseignements que Statistique Canada est autorisé à divulguer

Si la *Loi sur la statistique* garantit la confidentialité des réponses aux enquêtes, elle offre aussi le moyen de réduire le fardeau que constituent pour le répondant les sondages en double. En effet, la Loi permet à Statistique Canada de conclure des accords sur le partage de données avec d'autres ministères fédéraux et provinciaux et des entreprises. Si un accord de partage de données s'applique à une enquête particulière, les répondants en sont informés au moment de la collecte.

Les répondants peuvent aussi permettre la divulgation des renseignements les concernant en signant une renonciation qui permet de diffuser des renseignements sur un sujet particulier. La diffusion d'informations identifiables est régie par une politique et doit dans chaque cas être autorisée par le statisticien en chef du Canada.

Exercices

1. Dressez une liste de certaines organisations qui pourraient détenir des renseignements personnels à votre sujet.
    a. Énumérez les utilisations possibles de ces renseignements.
    b. À votre avis, lesquelles de ces utilisations seraient légitimes?
    c. Y a-t-il des utilisations qui, selon vous, ne seraient pas légitimes? Si oui, pourquoi?

2. Votre école détient-t-elle des données informatisées à votre sujet? Si oui, quelles sortes de données et quelle information pouvez-vous consulter? Une partie de cette information est-elle de nature confidentielle? Si oui, pourquoi?

3. Imaginez que vous êtes chargé de la sécurité informatique dans un organisme gouvernemental ou une entreprise. Que feriez-vous pour interdire l'accès sans autorisation aux renseignements personnels stockés dans les bases de données de l'organisme?

Ordinateurs et données

Il y a longtemps que les organisations professionnelles, les institutions et les entreprises utilisent des ordinateurs pour traiter des données et fournir de l'information. Bien que les ordinateurs et les systèmes informatiques soient devenus de plus en plus complexes et perfectionnés, leurs caractéristiques de base demeurent inchangées.

Essentiellement, les systèmes informatiques reposent sur une combinaison de matériel, de logiciels et d'utilisateurs (les gens). Chacun de ces éléments est essentiel au fonctionnement du système.

L'ordinateur électronique a fait son apparition dans le milieu des affaires au Canada et dans le monde dans les années 1960. Il a été adopté de façon graduelle parce que les premiers modèles étaient peu performants et entraînaient des coûts d'achat et des frais de maintenance élevés, et parce qu'il n'y avait que très peu de compétences spécialisées. Il y a vingt ans, environ 10 % du personnel et des étudiants dans une université typique utilisaient des ordinateurs — aujourd'hui, ce taux atteint près de 100 %. De plus, à l'époque, les compétences informatiques en gestion n'étaient pas enseignées à l'université et les applications logicielles de gestion n'étaient pas disponibles sur le marché — elles devaient être élaborées à l'interne par les programmeurs. Toutefois, les programmeurs compétents étaient rares et leurs services dispendieux.

Dans l'histoire de l'informatique, on parle souvent de « générations ».

- Les ordinateurs de la première génération (de 1940 à 1950) étaient principalement composés de circuits câblés de tubes à vide et utilisaient des cartes perforées comme principal support de stockage.
- Les ordinateurs de la deuxième génération (de 1950 à 1964) utilisaient des transistors. Le tube à vide peu efficace avait été remplacé par un composant beaucoup plus petit et plus fiable.
- Les ordinateurs de la troisième génération (de 1964 à 1972) utilisaient des circuits intégrés, invention qui a abouti à l'utilisation répandue des ordinateurs aujourd'hui. Les scientifiques ont trouvé une façon de réduire la taille des transistors, ce qui a permis de produire des ordinateurs plus petits.
- Les ordinateurs de la quatrième génération (de 1972 à aujourd'hui) utilisent des puces microprocesseurs qui constituent un circuit intégré à grande échelle contenant des milliers de transistors. Les transistors sur une puce microprocesseur peuvent exécuter toutes les fonctions de l'unité centrale de l'ordinateur. Le microprocesseur a permis de créer le premier ordinateur personnel.

Histoire des ordinateurs

Histoire des ordinateurs et de leur utilisation dans le secteur de l'éducation

- 500

L'abaque, la première machine à calculer au monde, aurait vu le jour il y a 2 500 ans dans la vallée du Tigre et de l'Euphrate. Initialement, l'abaque n'était qu'une tablette de pierre ou d'argile utilisée pour compter à l'aide de cailloux. Les cailloux étaient placés dans des rainures découpées dans la tablette et pouvaient être déplacés d'un côté à l'autre pour représenter des nombres. Cela facilitait le calcul, y compris des opérations simples comme l'addition et la soustraction.

1300

Au XIIIe siècle, la Chine a pensé à enfiler des grains ou des cailloux percés de part en part sur des ficelles ou des fils de métal fixés à un cadre. Cette découverte pose le fondement menant à la création de l'abaque moderne.

1600

L'écossais John Napier, inventeur du logarithme et du signe décimal, invente un dispositif manuel d'aide aux opérations de multiplication et de division, connu sous le nom de « réglettes de Neper ».

1622

L'Anglais William Oughtred invente la règle à calcul. Contrairement à celles qui suivront, la sienne est circulaire.

1623

L'Allemand Wilhelm Schikard construit une machine à calculer, connu sous l'appellation « horloge à calculer ». L'horloge est capable d'additionner et de soustraire des nombres composés allant jusqu'à six chiffres. La machine et les plans sont perdus, redécouverts en 1935, puis perdus et redécouverts de nouveau en 1956. La machine est reconstruite en 1960, et il paraît qu'elle fonctionnait.

1642

Le Français Blaise Pascal construit la Pascalline — le premier ordinateur numérique pouvant effectuer des additions. On pouvait y entrer jusqu'à 8 chiffres en bougeant des roues. Pascal construit et vend au moins une douzaine de ces appareils, dont certains existent encore de nos jours.

1668

L'Anglais Sir Samuel Morland construit une machine à additionner non décimale adaptée à la monnaie anglaise.

1673

L'Allemand Gottfried Leibniz, co-inventeur du calcul intégral, conçoit la calculatrice à étages, capable d'effectuer des multiplications allant jusqu'à 12 chiffres. Elle permet non seulement d'effectuer des additions et des soustractions, mais aussi des multiplications et des divisions, ainsi que d'extraire les racines carrées. Elle fut trouvée dans un grenier en 1879.

1775

Charles Earl Stanhope invente une calculatrice capable d'effectuer des multiplications.

1801

Joseph-Marie Jacuard invente un métier à tisser automatique à cartes perforées.

1820

Le Français Charles Xavier Thomas de Colmar invente l'arithmomètre, la première calculatrice capable d'effectuer avec succès des additions, des soustractions, des multiplications et des divisions.

1822

Charles Babbage met au point le premier ordinateur mécanique.

1832

Charles Babbage produit un prototype de la première calculatrice mécanique automatique. Sa « machine à différences » doit permettre de calculer et d'imprimer des tableaux mathématiques. Un septième seulement de cette machine est assemblé par l'ingénieur de Babbage, Joseph Clement. La plupart des 12 000 pièces fabriquées sont mises au rebut et fondues.

1833

Babbage entreprend la construction d'une machine à calculer analytique comprenant des éléments informatiques comme des unités d'entrée et de sortie et des systèmes mémoire.

1853

L'observatoire Dudley à Albany, New York, achète la première calculatrice, construite par le Suédois, George Scheutz, et son fils Edvard, pour produire une série de tableaux astronomiques. La machine de Scheutz est conçue sur la base du modèle de Babbage.

1854

Le mathématicien britannique George Boole conçoit l'algèbre binaire. Son travail est le fondement de la commutation binaire essentielle à l'informatique moderne.

1878

L'Espagnol Ramon Verea, qui habite à New York, invente une calculatrice munie d'une table de multiplication interne.

1886

Herman Hollerith, du Bureau of the Census des États-Unis, invente une machine à cartes perforées servant à compiler et à totaliser des données. Dorr E. Felt, de Chicago, conçoit la première calculatrice permettant d'entrer des nombres au moyen de touches à enfoncer au lieu d'un système à roues. Sa calculatrice est connue sous le nom de « comptomètre ».

1889

Felt invente la première machine à calculer imprimante.

1896

Hollerith fonde la Tabulating Machine Company qui devient éventuellement le International Business Machines (IBM) Corporation.

1931

À l'université Cambridge, E. Wynn-Williams utilise des thyratrons pour construire un compteur numérique binaire aux fins d'expériences en physique.

1935

International Business Machines lance l'IBM 601, une machine à cartes perforées dotée d'une unité arithmétique à relais, capable de faire une multiplication par seconde.

1937

George Stibitz de Bell Telephone Laboratories construit un additionneur binaire d'un bit à relais. Il s'agit de l'un des premiers ordinateurs binaires. Alan Turing invente sa machine universelle.

1939

John V. Atanasoff et Clifford Berry de l'université d'Iowa construisent un prototype d'additionneur de 16 bits. C'est la première machine à calculer utilisant des tubes à vide. Bell Telephone Laboratories invente la calculatrice à nombre complexe.

1941

Atanasoff et Berry mettent au point une calculatrice spéciale servant à résoudre les problèmes d'équations linéaires simultanées. Elle est connue plus tard sous le nom d'ABC (Atanasoff-Berry Computer). Cet ordinateur est doté d'une mémoire de 60 mots de 50 bits sous forme de condensateurs montés sur deux tambours tournants. Sa deuxième mémoire perfore des trous dans des cartes.

1943

Le Britannique Thomas Flowers construit le premier ordinateur électronique programmable, contenant 2 400 tubes à vide pour les opérations logiques. Cet ordinateur de déchiffrement, portant le nom Colossus Mark I, peut traduire 5 000 caractères par seconde et utilise une bande perforée pour l'entrée de données. Il est créé pour déchiffrer les messages produits par le dispositif de codage allemand Enigma.

1945

On trouve un « bogue » dans un relais informatique et le mot « déboguer » est créé.

1946

Les premiers ordinateurs à tubes à vide sont construits à l'Université de la Pennsylvanie. L'ENIAC (Electrical Numerical Integrator and Calculator) comprend 18 000 tubes à vide et occupe 1 800 pieds carrés. C'est le premier « véritable ordinateur » (ç.-à-d. le premier ordinateur numérique entièrement électronique à usage général).

1947

William B. Shockley, John Bardeen et Walter H. Brattain, scientifiques à l'emploi de Bell Laboratories aux États-Unis, inventent le transistor.

1951

On commence à voir les effets du baby boom dans l'augmentation de la taille des classes, mais les écoles ont peu recours à la technologie électronique. L'ordinateur Univac de la première génération est livré au Bureau of the Census des États-Unis. Whirlwind, le premier ordinateur en temps réel, est construit pour le Système de défense aérienne des États-Unis.

1952

Univac est utilisé pour prévoir les résultats de l'élection présidentielle de 1952. Personne ne croit sa prévision, fondée sur 1 % des voix, qu'Eisenhower remportera l'élection haut la main. Pourtant, c'est bien ce qui est arrivé.

1955

IBM vend son premier ordinateur commercial.

1957

On entreprend la construction au M.I.T. du premier ordinateur transistorisé, le TX-O (Transistorized Experimental Computer).

1958

Les gros ordinateurs hôtes ne sont pas généralement acceptés dans les écoles, qui continuent de transmettre l'information aux élèves en salle de classe par l'entremise de l'enseignant en tant que gestionnaire. Jack St. Claire Kilby, qui travaille chez Texas Instruments, invente le circuit intégré. Deux importants langages de programmation sont créés vers la fin des années 1950 — COBOL et LISP.

1959

Les ordinateurs transistorisés entrent en usage. Des ordinateurs plus petits utilisant des transistors et des circuits imprimés sont construits de 1959 à 1964. Ils sont considérés comme des ordinateurs de la deuxième génération.

1960

Mme Grace Murray Hopper, professeure de mathématiques achève l'élaboration du langage COBOL.

1962

Les lignes aériennes commencent à utiliser un système de réservations informatisé.

1963

Aux États-Unis, le Vocational Education Act affecte de nouveaux fonds au soutien de la technologie dans les écoles. Toutefois, les méthodes de traitement par lots utilisées par les gros ordinateurs et les mini-ordinateurs ne cadrent pas avec celles fondées sur l'enseignant en tant que gestionnaire de l'apprentissage employées dans la plupart des écoles. BASIC, un langage de programmation simple évolué, est conçu et utilisé principalement dans les universités pour former des programmeurs. Les ordinateurs de la gamme IBM 360 sont créés. La plupart des ordinateurs utilisent encore des unités d'entrée à cartes perforées. L'imprimante ligne par ligne demeure la principale unité de sortie. Digital Equipment construit le premier micro-ordinateur, appelé le PDP-9.

1964

Les ordinateurs construits entre 1964 et 1971 sont considérés comme des ordinateurs de la « troisième génération ». Ils comprennent les premiers circuits intégrés, qui permettent de construire des machines encore plus petites. IBM lance le langage de programmation PL/1. L'IBM 360 est lancé. La souris et les « fenêtres » sont inventées, de même que le langage de programmation BASIC. Gordon Moore (p.-d.g. d'Intel de 1979 à 1987) prédit que le nombre de transistors que l'industrie pourra placer sur une puce électronique doublera chaque année, prédiction qui plus tard sera connu sous le nom de « Moore's Law ».

1965

Des gros ordinateurs et des mini-ordinateurs sont installés dans certaines écoles. Ils sont utilisés principalement à des fins administratives ou d'orientation. On met au point le premier superordinateur, le Control Data CD6600.

1967

Les langages de programmation évolués comme FORTRAN sont enseignés dans les universités. Les programmes de formation professionnelle offerts dans les écoles commencent à comprendre l'entretien des ordinateurs.

1968

Création d'Intel Corporation.

1969

Le ministère de la Défense des États-Unis met sur pied le réseau ARPAnet (Advanced Research Projects Agency network) en vue des recherches sur le réseautage. Il pose les fondements du futur Internet. Le Net initial est un petit réseau de superordinateurs servant à promouvoir le partage de recherches parmi les universités. On assiste au Branchement des premiers ordinateurs hôtes.

1970

Les gros ordinateurs et les mini-ordinateurs sont utilisés dans certaines écoles, mais non de façon répandue dans l'enseignement. Intel lance la première puce RAM, la 1103, d'une capacité d'un kilo-octet ou 1 024 octets.

1971

Intel construit son premier microprocesseur, le 4004, capable d'environ 60 000 interactions par seconde (0,06 million d'instructions par seconde) et ayant une fréquence de base de 108 KHz. Les premiers micro-ordinateurs (ordinateurs personnels) voient le jour. Utilisation généralisée de gros ordinateurs et de mini-ordinateurs dans l'entreprise. Quelques nouvelles compagnies de logiciels créent des programmes éducatifs pour les gros ordinateurs et les mini-ordinateurs. Les ingénieurs d'IBM, sous la direction d'Alan Shugart, inventent le disque souple. On termine l'élaboration du langage de programmation Pascal (qui tient son nom du célèbre mathématicien).

1972

Les ordinateurs construits après 1972, dits de la « quatrième génération », comprennent des circuits intégrés à grande échelle (c.-à-d. des microprocesseurs), une puce comportant 500 composants ou plus. Création du langage de programmation C et, plus tard, du C++. La première calculatrice scientifique de poche (le HP-35), fabriquée par Hewlett-Packard, rend la règle à calculer obsolète. Intel lance le processeur 8008. La compagnie Automatic Electronic Systems du Canada lance le premier appareil de traitement de texte programmable avec écran au monde, le AES 90. Le système utilise des disques magnétiques pour le stockage des données et un microprocesseur spécialement conçu à cet effet.

1974

L'ordinateur Apple I est vendu sous forme d'appareil en trousse. Création du CLIP-4, premier ordinateur doté d'une architecture parallèle. Le MITS Altair 880, le Scelbi et le Mark-8 sont lancés. Ce sont les premiers ordinateurs personnels. Telenet voit le jour et la première version commerciale d'ARPAnet est mise en marché.

1975

Un certain nombre d'ordinateurs personnels Apple I sont offerts gratuitement aux écoles. Certaines écoles installent un gros ordinateur et des mini-ordinateurs, mais la plupart refusent de songer à se doter d'ordinateurs personnels. Bill Gates et Paul Allen mettent en oeuvre le langage BASIC pour la première fois. IBM lance la première imprimante au laser. Bill Gates et Paul Allen fondent la compagnie Micro-soft. (le tiret sera supprimé ultérieurement).

1977

Mise en marché d'Apple II.

1979

Selon les estimations, quinze millions d'ordinateurs personnels sont utilisés dans le monde entier. On assiste à la création de chiffriers pour ordinateurs personnels. Les gros ordinateurs et les mini-ordinateurs sont d'usage répandu. Le jeu d'arcade vidéo appelé « Space Invaders » lance la manie des jeux d'arcade. Honeywell crée le langage de programmation Ada, en souvenir d'Augusta Ada Byron, l'une des premières informaticiennes au monde et, chose étonnante, la fille de Lord Byron, le poète romantique bien connu.

1980

Le TI-99, fabriqué par Texas Instruments et doté d'un écran de télévision, devient l'ordinateur personnel le plus populaire au monde. L'élaboration de MS-DOS/PC-DOS commence. Le Sinclair ZX80 est lancé.

1981

IBM construit et met en vente un ordinateur personnel, premier fabricant de gros ordinateurs à le faire. Création des premiers programmes d'exercices scolaires pour ordinateurs personnels. Le Xerox Start System et le premier système WIMP (Windows, Icons, Menus, Pointing) sont lancés. ARPAnet compte 213 ordinateurs hôtes et s'élargit rapidement. Microsoft met en marché MS-DOS, version 1.0.

1982

Le protocole TCP/IP est adopté et le terme « Internet » est employé pour la première fois pour décrire l'ensemble de réseaux interconnectés qui utilisent le TCP/IP. Mise en marché du Commodore 64. Compaq met en marché son ordinateur Compaq Portable, qui est compatible avec l'OP d'IBM. IBM lance les unités de disques souples de 320 K à double face.

1983

Les clones de l'OP d'IBM prolifèrent. Sperry Corporation devient le deuxième fabricant de gros ordinateurs à produire un ordinateur compatible avec l'OP d'IBM (créé par Mitsubishi au Japon). L'ordinateur Apple II est généralement accepté dans le domaine de l'éducation parce qu'il répond au modèle d'enseignement axé sur l'enseignant comme gestionnaire. Création de programmes de simulation simples pour les ordinateurs personnels. IBM met en marché le PC junior.

1984

Il y a encore relativement peu d'ordinateurs dans les salles de classe. Création et mise en marché de l'ordinateur Macintosh Apple. Les fabricants de progiciels commerciaux créent des tutoriels et des jeux d'apprentissage informatiques. Le serveur DNS (serveur de nom de domaine) est adopté pour l'Internet, auquel sont reliés environ 1 000 ordinateurs hôtes. William Gibson invente le mot « cyberspace » (cyberespace) dans son roman Neuromancer.

1985

Microsoft met en marché Windows. Lotus, Intel et Microsoft adoptent la LIM EMS (norme de gestion de la mémoire paginée).

1987

Plus de 10 000 ordinateurs hôtes sont connectés à Internet.

1988

Création de l'ordinateur portatif et de la première puce optique. IBM met en marché pour la première fois les disques WORM (Write Once Read Many) ou disques monosession.

1989

Le World Wide Web, inventé par Tim Berners-Lee, vise à combler le besoin d'échange d'information à l'échelle mondiale. La carte de Sound Blaster est mise en marché.

1990

Création de l'ordinateur personnel multimédia. Les écoles utilisent des vidéodisques. Usage répandu des logiciels d'auteur multimédia orientés objet. Les simulations, les bases de données éducatives et d'autres types de programmes d'EAO (enseignement assisté par ordinateur) sont fournis sur disques CD-ROM, dont plusieurs incluent l'animation et le son. ARPAnet est mis hors service - le nombre d'ordinateurs hôtes dépasse les 300 000.

1991

Le Finois Linus Torvalds crée Linux, variante du système d'exploitation UNIX. Le monopole d'Intel est mis à l'épreuve avec le lancement par AMD du microprocesseur Am386.

1992

Les écoles utilisent des serveurs Gopher pour fournir aux étudiants de l'information en direct.

1993

Les fournisseurs de services commerciaux sont autorisés à vendre l'accès à Internet à des particuliers. Le Pentium. Mosaic, premier navigateur Web graphique, est mis en marché. Adobe introduit le format PDF (Portable Document File). AMD lance le microprocesseur de la gamme Am486 pour concurrencer le 80486 d'Intel.

1994

La vidéo numérique, la réalité virtuelle et les systèmes tridimensionnels retiennent beaucoup l'attention, mais les ventes d'ordinateurs personnels multimédia sont inférieures à celles d'ordinateurs personnels de gestion. Popularité croissante dans les écoles des systèmes d'auteur orientés objet comme HyperCard, Hyperstudio et Authorware. Netscape 1.0 est créé comme fureteur de rechange à NCSA Mosaic. La première technologie sans fil (Bluetooth) fait son apparition sur le marché. Yahoo lance son service de recherche Internet. Le World Wide Web compte environ 2 000 serveurs.

1995

Le recours à Internet et au World Wide Web s'accroît au fur et à mesure que les entreprises, les écoles et les particuliers créent des pages Web. La plus grande partie de l'enseignement assisté par ordinateur se donne à l'aide de disques CD-ROM, dont la popularité ne cesse de s'accroître. Mise en marché de Windows 95 et de Pentium Pro. Netscape annonce le lancement de Javascript. George Moore révise sa loi, en préconisant que les transistors placés sur les circuits intégrés doubleront désormais tous les deux ans. La suprématie d'Intel dans le monde des micro-ordinateurs sous Windows continue de s'effriter avec le lancement par AMD du microprocesseur K-5 qui, compatible broche à broche et à un prix inférieur, représente une alternative sérieuse au Pentium d'Intel.

1996

Internet retient l'attention à mesure que les entreprises commencent à fournir des services et à faire de la publicité sur les pages Web. On assiste à la création de nouveaux outils graphiques et multimédias pour fournir de l'information et dispenser de l'enseignement sur Internet. De nombreuses écoles procèdent à un recâblage pour l'accès à Internet. Quelques écoles installent des serveurs Web et donnent à leur corps enseignant le moyen de créer des pages Web éducatives. On procède à la mise en marché de Netscape Navigator 2.0. Le nombre d'ordinateurs hôtes sur Internet atteint presque les 10 000 000. Microsoft lance la première version d'Internet Explorer, son navigateur Web breveté.

1997-1998

Internet continue de prendre de l'expansion grâce à de nouvelles utilisations et applications. La reconnaisssance de la parole est de plus en plus répandue. On s'attend à une demande accrue de logiciels éducatifs étant donné les CD-ROM de capacité beaucoup plus importante. Intel lance le Pentium MMX pour les jeux et les activités multimédias. Intel met en marché le processeur Pentium II. Microsoft lance Windows 98 et AMD procède au lancement du microprocesseur K-6. Palm Computing Markets met en marché le premier PDA (Personal Digital Assistant), le Palm Pilot. On assiste aussi à l'arrivée de la technologie E-book. L'arrivée des programmes téléchargeables comme SETI@Home ouvre la voie au traitement de l'information par Internet sur une grande échelle.

1999

On procède au lancement de Linux Kernel 2.2.0. On estime qu'environ 10 millions de personnes utilisent Linux. AMD met en marché le K6-III, version 400 MHz, qui contient environ 23 millions de transistors. Dans certains tests, il donne un meilleur rendement que l'Intel P-III. Intel procède au lancement du Pentium de la gamme P-III. Un grand nombre de sites de commerce électronique font leur apparition sur Internet. Les gouvernements et les entreprises du monde entier se mobilisent en vue de l'arrivée de l'an 2000 (Y2K). Bien que les nouveaux ordinateurs soient compatibles à l'an 2000, on s'inquiète du fait qu'il existe toujours un grand nombre d'ordinateurs vulnérables. AMD lance sa nouvelle puce Athlon brevetée. Grâce à sa vitesse de 1 GHz, elle surpasse tous les microprocesseurs Pentium offerts par Intel.

2000

Les craintes liées au passage de l'an 2000 s'avèrent totalement non fondées, l'arrivée du nouveau millénaire se déroulant sans aucune interruption dans les réseaux informatiques mondiaux. Microsoft lance Windows Millennium (pour les OP) et Windows 2000 (pour les réseaux locaux). Les titres des entreprises dot-com plongent à la bourse, et plusieurs d'entre elles déclarent faillite. Le président de Microsoft, Bill Gates, démissionne comme p.-d.g. de sa propre entreprise pour se consacrer à l'élaboration de logiciels. IBM met en marché une version de suivi de Deep Blue, appelée Blue Gene : elle fonctionne au rythme d'un quadrillion d'opérations par seconde (un petaflop), soit 1 000 fois plus vite que Deep Blue. Blue Gene servira à la modélisation de protéines humaines. Des cyberattaques obligent certains sites Web à fermer.

2001

Détection du premier virus Linux. Gordon Moore déclare que la loi qui porte son nom devrait être modifiée, le nombre de transistors sur un circuit intégré devant doubler non pas tous les deux ans mais tous les quatre ou cinq ans, à compter de 2010-2020. En 1965, il avait prédit que leur nombre doublerait chaque année puis, en 1995, il avait dit tous les deux ans. Le fait que cette prévision ait tenu pendant 36 ans est extraordinaire, étant donné la rapidité de l'évolution dans ce domaine. Intel lance le Pentium IV, alors que Microsoft met en marché le Windows XP (pour les OP et les réseaux locaux).

2002

La technologie sans fil est de plus en plus répandu. De nouveaux appareils portatifs sont disponibles sur le marché, qui comprend entre autres les modems de communications sans fils, les téléphones cellulaires en mode mixte, les navigateurs Web, les ordinateurs de poche, les récepteurs GPS (Global Positioning System) et les systèmes d'exploitation et d'interfaces graphiques de plus en plus sophistiqués comme le PC Tablet.

Historique des ordinateurs à Statistique Canada

Introduction

La *Loi du recensement et des statistiques* de 1905 marque le début de l'ère statistique moderne par la création d'un bureau permanent du recensement et de la statistique, connu aujourd'hui sous le nom de Statistique Canada.

Historique des ordinateurs à Statistique Canada

1911

Le Recensement se compose de 13 questionnaires comportant 522 questions. Les machines utilisées pour le traitement des données des recensements de 1911 à 1942 ont été fabriquées sur mesure par A.E. Thornton, l'un des préposés à l'exploitation mécanographique du Bureau.

1918

Le gouverneur général en conseil approuve l'utilisation d'appareils mécaniques pour toutes les compilations statistiques du gouvernement. Les cartes sont perforées aux ministères mêmes, puis transférées au nouveau Bureau fédéral de la statistique où elles sont triées et mises en tableaux.

1919

Le ministère des Finances utilise les machines du Bureau pour compiler les données relatives à l'impôt sur le revenu.

1921

Bien que la compilation et la mise en tableaux soient presqu'entièrement mécanisées, un nouveau classicompteur inventé par Fernand Bélisle, préposé à l'exploitation mécanographique au Bureau, accroît grandement la productivité. Trois de ces machines sont construites, ce qui permet de totaliser 50 fois plus de renseignements que les anciens appareils et de réaliser d'importantes économies au titre de la main-d'oeuvre.

1932

R.-H.-Coats propose la création d'un organisme de recherche sociale et économique. Selon lui, un tel organisme « ... répondrait à moyen terme à ce qui est sans aucun doute devenu l'un des besoins pressants du gouvernement—se tenir au courant de l'évolution de la pensée économique, toujours disposer des données requises aux fins de la formulation de ses grandes politiques économiques et être doté des machines spécialement conçues à cette fin. » [Traduction]

1949

La Mechanical Tabulation Division (MT) compte un effectif de 202 employés utilisant 173 appareils mécanographiques et 27 machines à additionner ou comptomètres. La division se compose d'une sous-section de la perforation desservant plusieurs divisions clientes, d'un certain nombre de sous-sections de traitement chargées de répondre aux besoins de clients particuliers et d'une sous-section de machines auxiliaires assurant le soutien des unités de traitement. En outre, des machines à calculer et à additionner et des dispositifs en panneau perforé servent à effectuer des compilations plus simples, directement à partir des renseignements statistiques fournis.

1950

Le U.S. Bureau of the Census (bureau du recensement des États-Unis) utilise l'ordinateur UNIVAC pour traiter les données du Recensement de 1950. En 1951, un rapport précise qu'il faudrait au moins 650 opérateurs et 17 perforatrices de cartes pendant la semaine de pointe du dépouillement du recensement pour produire un million de cartes perforées complètement vérifiées et prêtes à la mise en tableaux. Le Bureau, souhaitant réaliser des gains de productivité et s'intéressant par ailleurs aux innovations techniques, achète une machine à calculer électronique.

1960

Herbert Marshall, le statisticien fédéral, annonce que le Bureau s'apprête à installer un ordinateur électronique. En 1960, un ordinateur IBM 705 est acheté pour dépouiller les données du Recensement de 1961. Il s'agit de l'un des plus puissants ordinateurs au Canada, si bien que deux gros climatiseurs ont dû être installés pour dissiper la chaleur générée par ses 10 000 tubes à vide.

1962-1970

En 1962, Statistique Canada dispose d'un IBM 705 et d'un IBM 1401 de capacité accrue. En 1963, le Bureau acquiert un IBM 101 et, en 1964, un deuxième IBM 1401. En 1966, le premier IBM 1401 est remplacé par un IBM 360/30. Le Recensement de 1966 est dépouilllé de la même manière que celui de 1961. Le Système de stockage et d'extraction des données codées suivant une grille géographique (GRDSR) et la banque de données chronologiques du Système canadien d'informations socio-économiques (CANSIM) constituent les principaux progrès réalisés au titre des systèmes utilisés à cette époque. En 1969, l'IBM 360/30 est remplacé par un IBM 360/65.

1971

L'année 1971 marque le centenaire du premier recensement du Dominion du Canada. Le dépouillement des questionnaires sur la population et le logement se fait à partir de 12 caméras à avancement de bande automatique fournies par le Bureau of the Census des États-Unis. Les données sont ensuite transférées sur bande magnétique au moyen d'un nouveau lecteur de documents, l'Appareil de lecture optique pour transfert automatique (FOSDIC). Le traitement informatique se fait au moyen d'un IBM 360/65, mis à niveau de manière à doubler la mémoire centrale (1 024 000 octets). De nombreux tableaux sont produits sous forme d'imprimés informatiques, puis microfilmés pour la diffusion aux utilisateurs. Une bande sommaire est produite également pour répondre à la demande de données agrégées sous forme lisible à la machine.

1972

L'IBM 360/65 est remplacé par un IBM S/370-165 d'une capacité de 1 536 000 octets.

1973

L'IBM S/370-165 a maintenant une capacité de 2 000 000 d'octets et une mémoire à disques magnétiques remarquable de 3,6 milliards d'octets. En plus de l'ordinateur central, le Bureau fait l'acquisition de micro-ordinateurs et de mini-ordinateurs autonomes.

1974

L'ordinateur central IBM S/370-168 a une capacité de 3 millions d'octets et une mémoire à disque de 5,2 milliards d'octets.

1976

Statistique Canada construit un nouveau centre informatique dans l'immeuble Principal. À cette époque, l'ordinateur central est l'Amdahl 470/V6, muni d'une mémoire à disques de 10 milliards d'octets.

1981

La révolution causée par l'arrivée des ordinateurs personnels débute modestement à Statistique Canada, avec l'achat de quelques appareils Xerox Star. Le Bureau achète ensuite d'autres ordinateurs personnels compatibles IBM, dont le nombre ne cesse d'augmenter par la suite.

1982

Les installations informatiques sont centralisées et on passe à un appareil Amdahl/V8.

1985

L'ordinateur central Amdahl du Bureau passe à la version 5870/5880.

1987-1988

L'ordinateur central est remplacé par un IBM 3090, afin de permettre au Bureau de traiter entre 5 000 à 6 000 travaux par jour—traitement de données d'enquête ou mise à jour et interrogation de bases de données. Le Bureau achète 200 ordinateurs personnels, portant leur nombre total à près de 1 000.

1989

Statistique Canada acquiert la première mémoire de masse à cartouches qui assure automatiquement le stockage et la récupération des cartouches dans une zone.

1990

L'ordinateur central est remplacé par un Amdahl 5990. Dans les bureaux régionaux, des réseaux locaux sont mis en place dans le cadre d'un programme pluriannuel de mise à niveau des installations de traitement électronique des données.

1991

L'ordinateur central est remplacé par l'IBM 3090/600S, qui est 140 fois plus puissant que le système acheté en 1969. La mémoire à disques est portée à 1,2 billion d'octets. On compte aussi plus de 4 000 ordinateurs personnels, soit un pour chaque employé.

1993

Pour améliorer la qualité des données et réduire le fardeau de réponse, les intervieweurs des bureaux régionaux font l'essai sur place d'ordinateurs portatifs.

1996

On remplace l'ordinateur central par la version IBM 9672-R44, comportant une mémoire de 2 gigaoctets et des canaux à fibres optiques (ESCON). La mémoire à disques est portée à 1,2 billion d'octets.

1999

On annonce le lancement des Services du système Unix, qui offrent aux utilisateurs la possibilité d'exécuter les charges de travail suivant la méthode traditionnelle des lots et de l'option de temps partagé (TSO) ou les charges de travail Unix depuis l'ordinateur central.

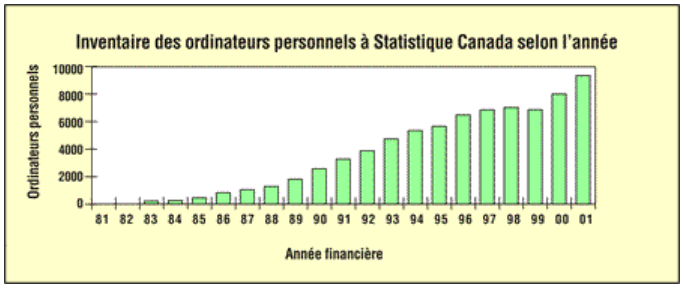
2001

Le sous-système à disque de l'ordinateur central est mis à niveau afin de fournir une capacité de stockage de 2,5 billions d'octets sur disque protégé par le réseau redondant de disques indépendants (RAID).

2002

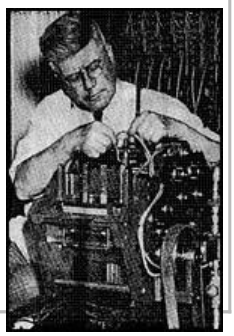
Le sous-système à bande magnétique de l'ordinateur central est mis à niveau afin d'adopter la technologie de la bande virtuelle (Virtual Tape Technology). Cette technologie permet à toutes les données de la bande magnétique de l'ordinateur central d'être versées dans une bande magnétique moderne à grande capacité sous contrôle robotique, éliminant ainsi la nécessité des étagères de rangement pour les 75 000 bandes magnétiques de l'ordinateur central du Bureau. Cette nouvelle technologie fournit aussi automatiquement une deuxième copie redondante de toutes les données sur bande magnétique afin d'éliminer les erreurs d'entrées-sorties une fois pour toutes.

Le graphique ci-dessous montre l'utilisation croissante d'ordinateurs à Statistique Canada, reflétant leur utilisation croissante dans le milieu des affaires.



Un regard sur notre passé

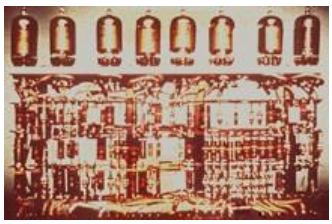
Un inventeur du Bureau conçoit des tabulatrices mécaniques



Fernand Bélisle, né dans les Cantons de l'Est, au Québec, en 1889, est devenu une figure légendaire au sein du BFS [Le Bureau fédéral de la statistique] à titre de génie de la mécanique. Le rapport administratif du statisticien fédéral [R.-H.-Coats] pour l'année 1941 attribue à Bélisle l'élaboration de la machine « Pantograph », utilisée depuis 1911 pour la perforation des cartes, ainsi que l'invention de la multiperforatrice électrique, utilisée avec une telle efficacité au Recensement de 1931, et du classicompteur à air comprimé qui « a permis de totaliser une quantité record de renseignements lors du Recensement de 1941 » (traduction libre). En faisant allusion à la tabulatrice à air comprimé, Coats affirmait à l'occasion de la Conférence sur les affaires canado-américaines de 1940 que le BFS « avait élaboré aux fins du recensement une machine qui fait l'envie du monde entier » (traduction libre). Exception faite de quelques interruptions de service en début de carrière, Bélisle a travaillé sans arrêt pour le compte du Bureau de 1911 jusqu'à sa retraite en 1950. Il est décédé en 1963.

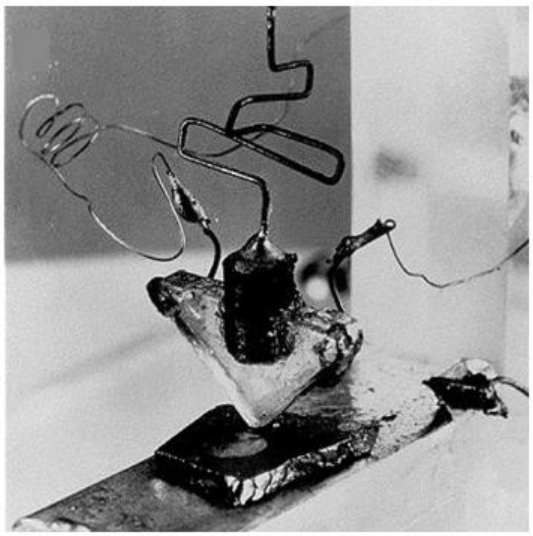
Un aperçu de l'ordinateur IBM 705 III en 1969

tube à vide



microprocesseur

transistor



Industrie de l'informatique

Installations informatiques commerciales

Habituellement, les grandes entreprises n'ont pas qu'un seul gros ordinateur, mais une série d'ordinateurs spécialisés qui sont réseautés. Une entreprise, selon sa taille, peut disposer d'un ou deux ordinateurs de très grande puissance, d'un certain nombre d'ordinateurs intermédiaires (Unix) de puissance faible, moyenne et grande, et de nombreux ordinateurs personnels (OP).

Dans la plupart des organisations, les ordinateurs sont montés en réseau. En général, un réseau qui relie les ordinateurs d'un endroit donné s'appelle un réseau local (RL). Des réseaux locaux reliés entre eux constituent un réseau étendu (RE), qui comprend des serveurs, des postes de travail, un système d'exploitation de réseau et une liaison de communication. Les serveurs sont des appareils ultrarapides qui gardent en mémoire les programmes et les données partagés par les utilisateurs du réseau. Grâce à un tel réseau, le personnel autorisé peut accéder aux données et utiliser les installations informatiques n'importe où au sein d'une entreprise décentralisée.

Jusqu'à il y a 15 ou 20 ans, les entreprises disposaient surtout de gros ordinateurs. De nos jours, ces ordinateurs sont principalement utilisés pour exploiter d'anciennes applications (souvent appelées « systèmes existants ») conçues il y a plusieurs années. Selon les exigences du système d'application, elles sont conçues pour être exploitées au moyen d'un gros ordinateur, d'ordinateurs intermédiaires (serveurs) ou d'ordinateurs personnels réseautés.

Actuellement, les utilisateurs dûment autorisés peuvent accéder aux fichiers de données au moyen de leur ordinateur, quelle que soit la plate-forme où se trouvent leurs données.

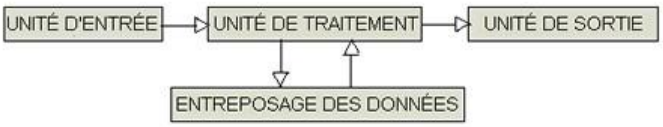
Internet et intranet

Internet a créé une véritable révolution dans le monde de l'informatique et des communications. L'invention du télégraphe, du téléphone, de la radio et de l'ordinateur ont ouvert la voie à cette intégration inégalée des capacités. Internet est à la fois un moyen de diffusion d'information à l'échelle mondiale et un moyen de collaboration et d'interaction entre les particuliers et leurs ordinateurs, partout dans le monde, et sans contraintes de temps.

Aujourd'hui, les entreprises, les établissements d'enseignement et les particuliers utilisent Internet pour transmettre des messages électroniques aux quatre coins du monde et pour mener des recherches. Toutefois, la plupart des entreprises trouvent Internet trop lent, peu fiable et trop peu sécuritaire pour en faire une utilisation généralisée à l'interne. Pour résoudre ce problème, elles ont mis sur pied leurs propres réseaux internes appelés intranets. L'intranet allie la commmodité d'Internet au rendement et à la sécurité d'un système maison. Un intranet peut être relié à Internet au moyen d'une passerelle protégée munie de pare-feu pour empêcher tout accès extérieur non autorisé aux systèmes internes.

Matériel informatique

Pour bien fonctionner, un système informatique doit comprendre les composants matériels suivants :



- une unité d'entrée qui permet à l'utilisateur d'entrer des données et d'exécuter des programmes informatiques;
- une unité centrale qui contrôle toutes les activités au sein du système;
- un disque dur ou des disquettes de stockage des données, des fichiers et des programmes;
- une unité de sortie qui permet de présenter les données ou l'information à l'utilisateur.

Unité d'entrée

Il existe différentes façons d'entrer des données dans un ordinateur :

- par scanner, clavier ou souris;
- un disque souple ou un CD-ROM;
- une bande magnétique, un photostyle ou un lecteur de code à barres, un microphone ou une caméra numérique;
- une ligne téléphonique ou de communication;
- le courrier électronique sur Internet et les bases de données Internet.

Unité centrale

L'unité centrale (UC) est le coeur du système informatique. Dans la plupart des ordinateurs modernes, elle se compose d'une ou deux puces de silicium qui sont suffisamment petites pour tenir dans une main, mais qui contiennent des millions de circuits logiques. Une UC peut exécuter des millions d'instructions par seconde.

L'UC est liée à une mémoire vive (RAM) qui doit être de taille suffisante pour garder tous les programmes et toutes les données exploitées à un moment donné. La mémoire vive peut contenir de quelques millions à quelques centaines de millions d'octets.

Stockage de données

L'ordinateur stocke les données dans des unités de disques. Aujourd'hui, un ordinateur personnel peut être muni d'une unité de disques de 2,0 gigaoctets (GO) ou 20 milliards d'octets, alors qu'un gros ordinateur peut être doté de plusieurs unités de disques pouvant contenir des dizaines ou des centaines de gigaoctets.

Les données peuvent être copiées sur bande magnétique ou CD-ROM aux fins de sauvegarde, de transfert entre ordinateurs ou de stockage à long terme. Un CD-ROM peut comprendre 700 mégaoctets, alors qu'une unité de bande magnétique peut comprendre entre 500 mégaoctets et 24 gigaoctets environ. Les données peuvent aussi être copiées sur des disquettes (disques souples), mais leur petite taille (1,44 mégaoctet) limite leur utilité pour les gros ensembles de données.

Unités de sortie

On entend par unité de sortie toute unité périphérique qui présente des sorties d'ordinateur. Ces unités comprennent :

- les moniteurs vidéo;
- divers types d'imprimantes;
- les disques et bandes magnétiques;
- les CD-ROM;
- les lignes téléphoniques et de transmission de données;
- les haut-parleurs stéréophoniques.

On peut aussi utiliser des ordinateurs pour commander des procédés industriels, pour contrôler des usines chimiques et pour verrouiller et déverrouiller des portes de sûreté. Les systèmes modernes de gestion des moteurs automobiles, les ascenseurs des tours à bureaux, les magnétoscopes à cassettes et de nombreux autres systèmes domestiques et industriels sont maintenant contrôlés par des systèmes informatiques miniatures.

Stockage et récupération

Une bonne partie de la puissance de l'ordinateur tient à sa capacité de stocker ainsi que de trier et de classer les données. Au cours des dernières années, les systèmes à disques sont devenus très fiables et bon marché; il est maintenant possible d'obtenir pour quelques centaines de dollars des systèmes à disques capables d'emmagasiner des milliards d'octets. Grâce à cette énorme baisse de prix ainsi qu'à la création de systèmes de base de données perfectionnés, l'ordinateur est devenu beaucoup plus utile. Il est maintenant possible de garder en mémoire sur ordinateur toute l'information de l'entreprise portant sur plusieurs années.

Logiciels

Essentiellement, il y a deux types de logiciels : le logiciel de base, qui contrôle l'exploitation de l'ordinateur, et le logiciel d'application qui permet à l'utilisateur d'exécuter des tâches utiles.

Les **logiciels de base** se divisent en deux catégories : le logiciel d'exploitation et les outils et logiciels utilitaires.

Un **logiciel d'exploitation** comprend habituellement deux niveaux :

- Au *niveau inférieur, le système d'entrée-sortie de base* (BIOS) contrôle les fonctions élémentaires, comme la mémoire vive à lecture et à écriture et la gestion des entrées et sorties d'unités périphériques telles que la souris, le clavier, l'imprimante et l'écran.
- Au *niveau supérieur, le système principal d'exploitation* (p. ex. Windows) fait fonction de plate-forme d'accueil des programmes. Il assure le lien qui permet à l'utilisateur de contrôler le fonctionnement de l'ordinateur et d'exploiter efficacement le logiciel d'application. Ainsi, il comprend un sous-système de fichiers comprenant des noms d'unités, des répertoires, des données, des fichiers et des index; en outre, il comporte des fonctions de gestion des fichiers (p. ex. création, copie, suppression).

DOS, UNIX, Mac OS et Windows sont des *systèmes d'exploitation typiques*. Mac OS et Windows sont dotés d'une interface graphique (GUI) conviviale qui permet de contrôler l'ordinateur au moyen de fenêtres, de menus, d'icônes et d'une souris. Dans le cas de DOS et d'UNIX, l'utilisateur doit taper des commandes précises dont il peut être difficile de se rappeler. Windows, Mac OS et UNIX peuvent exploiter de nombreux programmes en même temps. La multiprogrammation permet une utilisation plus efficace du matériel, tout en étant plus pratique pour l'utilisateur.

Les **outils et les logiciels utilitaires** sont généralement nécessaires aux fins de l'exploitation productive de l'ordinateur. Certains sont founis avec le système d'exploitation et d'autres, vendus séparément. Des exemples typiques sont les navigateurs Internet, les logiciels antivirus, les compilateurs de programmes, les programmes d'édition et les systèmes de sauvegarde de fichiers.

Les logiciels d'application se divisent aussi en deux catégories : les outils de productivité personnelle et les autres applications informatiques.

Outils de productivité personnelle

Ces outils sont des produits que l'on achète pour exécuter des tâches informatiques courantes comme le traitement de texte, l'analyse numérique, la manipulation et le stockage des données et la présentation des données. Quelques produits typiques sont décrits ci-dessous :

- **Traitement de texte :** Le logiciel de traitement de texte est conçu aux fins de la création de documents tels les lettres, les rapports, les articles de journaux et les livres. Ce type de logiciel a été l'une des premières applications créées pour les ordinateurs personnels, servant à rationaliser d'importants volumes de texte dactylographié. Le succès de ce type de logiciel tenait à ce qu'il permettait de corriger le texte sans avoir à retaper tout le document. MS Word, MS Works, Word Perfect et Word Pro sont des logiciels typiques de traitement de texte.
- **Chiffriers :** Les chiffriers servent à la comptabilité électronique et sont utilisés par les comptables pour organiser l'information commerciale. Ils sont très utiles lorsqu'il s'agit de manipuler des tableaux. Ils permettent de faire rapidement les opérations d'addition, de soustraction, de division, de

multiplication et de totalisation, ainsi que de recalculer automatiquement tous les résultats plus tard si de nouvelles données sont entrées. Les programmes de mise en forme et de création de graphiques facilitent l'analyse et la présentation. Des centaines de fonctions permettent de faire automatiquement des calculs statistiques, techniques, économiques et commerciaux, p. ex. les fonctions de calcul de l'intérêt composé et de l'écart-type. MS Excel, MS Works et Lotus 123 sont des exemples de chiffriers électroniques bien connus.

- **Bases de données :** Les progiciels de gestion de bases de données facilitent l'organisation et le stockage uniformes de données. Ils permettent de chercher, trier et présenter rapidement et systématiquement les données. Les bases de données peuvent être utilisées par des personnes qui n'ont aucune formation spéciale pour créer des listes d'envoi ou dresser des inventaires de magasins, de même que par des programmeurs professionnels pour produire des applications complexes destinées à aider à gérer une entreprise. MS Access, MS Works et Lotus Approach sont des exemples de progiciels de gestion de bases de données.
- **Progiciels de présentation :** Les progiciels de présentation sont utilisés pour illustrer les conférences et les exposés. Ces progiciels ont pratiquement remplacé les transparents de rétroprojection dessinés à la main ou dactylographiés. Dans de nombreux exposés et conférences aujourd'hui, le présentateur branche un ordinateur portatif directement sur un projecteur de diapositives. MS PowerPoint et Lotus Freelance sont des progiciels de présentation typiques.
- **Graphiques :** Les progiciels graphiques permettent aux utilisateurs de créer des dessins, de peindre des tableaux ou de mettre en relief des images balayées telles que des photographies et des illustrations. MS Paint est un progiciel graphique typique.
- **Éditique :** Les progiciels éditiques visent à améliorer l'apparence et la mise en page finale du texte et des graphiques à des fins de publication. Les graphiques ou illustrations peuvent provenir d'une picothèque. Ils peuvent aussi être créés au moyen de progiciels graphiques. MS Publisher, Harvard Graphics, Paint Shop Pro et PageMaker sont des progiciels d'éditique typiques.

Autres applications informatiques

Les outils de productivité personnelle sont utilisés à la maison, à l'école et au bureau. Ils sont relativement bon marché et faciles à installer sur de nombreux ordinateurs. Toutefois, les entreprises et les organisations achètent habituellement des ordinateurs pour automatiser les principales fonctions de gestion, ce qui normalement ne se fait pas au moyen de logiciels de productivité personnelle.

Certaines applications logicielles peuvent coûter des milliers de dollars à acheter ou à mettre au point, alors qu'un important système bancaire ou de réservation de billets d'avion peut coûter des millions. Le logiciel d'application peut être acheté sous forme de logiciel de série ou être conçu à une fin particulière. Un progiciel de comptabilité est un exemple typique d'une application de série, alors qu'un programme de traitement de contraventions pour stationnement interdit sera peut-être créé de toutes pièces.

Le système de navigation sur pilote automatique d'un avion est un exemple de logiciel qui reçoit de l'information (p. ex. le cap au compas et les lectures du système mondial de localisation [GPS]) et qui contrôle le système hydraulique actionnant le gouvernail et les ailerons, qui permettent de modifier la direction de l'avion.

Analyste de systèmes, programmeurs et utilisateurs

Les analystes de systèmes et les programmeurs élaborent des logiciels d'automatisation des tâches.

Analystes de systèmes

L'analyse de systèmes est le processus qui consiste à décomposer un problème de traitement de données en composantes fonctionnelles afin de déterminer la meilleure façon de procéder. Les analystes de systèmes conçoivent et modifient les systèmes en transformant les besoins des utilisateurs en un ensemble de spécifications fonctionnelles qui constitue le schéma directeur du système.

En consultation avec l'utilisateur, l'analyste de systèmes doit :

- définir les problèmes que présente le système d'une organisation;
- examiner la situation pour déterminer les besoins du nouveau système;
- élaborer des spécifications pratiques, efficaces et rentables, et qui font une utilisation optimale du matériel et du logiciel disponibles;
- communiquer les besoins du nouveau système à toutes les parties intéressées;
- aider à la mise en oeuvre du nouveau système.

Programmeurs

La programmation est le processus qui consiste à produire un ensemble d'instructions pour amener l'ordinateur à exécuter une activité donnée. Les programmeurs conçoivent des programmes à partir des résultats de l'analyse des systèmes en vue de résoudre les problèmes.

En consultation avec l'utilisateur, le programmeur doit :

- comprendre les problèmes et prévoir des solutions;
- concevoir des programmes en utilisant des diagrammes de cheminement des données et d'autres outils de conception;
- écrire les programmes nécessaires pour mettre en oeuvre la conception logique,
- mettre à l'essai les programmes et corriger tout problème éventuel;
- rédiger des documents détaillés sur les programmes et leur exploitation.

Utilisateurs

L'utilisateur est celui qui, en définitive, détermine si un système informatique répond aux besoins pour lesquels il a été conçu. Meilleur est le lien entre les composantes automatisées du système et l'utilisateur, plus il est probable que le système soit efficace. Les concepteurs modernes consultent beaucoup les utilisateurs afin de concevoir des systèmes qui répondent à leurs besoins et ils déploient beaucoup d'efforts et d'ingéniosité dans la conception de l'interface entre le système et les utilisateurs.

Exercices

1. Les systèmes informatiques sont une combinaison de trois composantes. Quelles sont ces composantes?

2. Nommez trois composants matériels d'un système informatique.

3. Donnez deux raisons qui expliquent pourquoi les grandes organisations créent leur propre système interne (intranet).

4. En quoi les chiffriers sont-ils utiles à une personne qui procède à l'analyse de données?

5. Pourquoi est-il important qu'un analyste de systèmes consulte l'utilisateur du système avant d'en concevoir un nouveau?

6. Dans quel pays le premier ordinateur électronique a-t-il été conçu?

7. Comment s'appelait la première machine à calculer?

8. Quelle est la différence entre « intranet » et Internet?

Réponses

1. Le matériel, les logiciels et les utilisateurs.

2. Les trois composants matériels sont les suivants :
    a. Les unités d'entrée : disquette, CD-ROM, bande magnétique, photostyle, lecteur de code à barres, microphone, caméra numérique, ligne téléphonique, autres lignes de communication.
    b. Les unités de stockage de données : disquette, CD-ROM, bande magnétique, unité de disques.
    c. Les unités de sortie : moniteur vidéo, imprimante, bande magnétique, disquette, CD-ROM, lignes de transmission de données, lignes téléphoniques, haut-parleurs stéréophoniques.

3. Pour allier la commodité d'Internet au rendement et à la sécurité d'un système maison.

4. Les chiffriers sont très utiles pour l'analyse des données, puisqu'ils permettent de manipuler automatiquement les données disposées sous forme de tableaux.

5. Il est important qu'un analyste des systèmes consulte l'utilisateur du système pour :
    a. cerner les problèmes que présente le système de l'organisation
    b. déterminer les besoins du nouveau système
    c. concevoir un nouveau système pratique, efficace et, rentable, et qui fait une utilisation optimale du matériel et du logiciel disponibles.

6. En Grande-Bretagne, par Thomas Flowers.

7. L'abaque.

8. Un réseau interne 'intranet' est un microcosme d'Internet mis sur pied dans une grande entreprise pour faciliter le travail. Les réseaux intranet sont plus fiables et mieux protégés qu'Internet. Ils sont aussi munis de « pare-feux » pour empêcher tout accès non autorisé de l'extérieur.

Bibliographie

A

AGRICULTURE ET AGROALIMENTAIRE CANADA. 2001. *Types de données*, Direction générale de la recherche. Adresse Internet (consulté le 22 janvier 2002).

ALLARD, Benoît, et Laurent ROY. 1998. « Atelier sur la conception de questionnaires : Statistique Canada, cours 410F », produit nº 10H0047XPF au catalogue de Statistique Canada, Centre d'information sur la conception des questionnaires, Ottawa.

ALOHA-SIDAWAY, Janice, et Margaret MCKINNON. 1999. « Évaluation de la valeur pédagogique des nouvelles technologies », *Une collection d'idées pour concevoir du matériel didactique efficace*, Hull, Patrimoine canadien.

ANDERSON, Jon E. 1999. « Chapter 3: Producing Data: Systematic Sampling », *Course Syllabus for Statistics 1601*, University of Minnesota. Adresse Internet (consulté le 4 mai 2001).

ATELIER D'INTERPRÉTATION DE DONNÉES. 1997. « Diagrammes pour publication, cours 04482 », calendrier du 3 novembre au 12 décembre 1997, produit nº 10H0053GPF au catalogue de Statistique Canada, Ottawa.

AUSTRALIAN BUREAU OF STATISTICS. 1990. *Australia: Working It Out! Core material for Australian studies*, Queanbeyan, AU, Australian Bureau of Statistics.

B

BLACK, Paul E. 1998. « The Role of Graphics », *Engineering Statistics Handbook*, version précommercialisation, National Institute of Standards and Technology. Adresse Internet (consulté le 7 août 2001).

BOCK, Rudolf K., et W. KRISCHER. 1998. « Stratified Sampling », *The Data Analysis Briefbook*, version 16. Adresse Internet (consulté le 4 mai 2001).

BRADY, Bennett M., et Joseph W. DUNCAN. 1978. *Statistical services in ten years' time*, Oxford, RU, Pergamon Press.

BROSTER, Eric James. 1974. *Glossary of Applied Management and Financial Statistics*, Essex, RU, Gower Press.

C

CAILLIAU, Robert. 1995. *A Little History of the World Wide Web*. Adresse Internet (consulté le 26 novembre 1999).

CASEY, Nancy, et Mike FELLOWS. 1998. « Graphs, Games, and the NCTM Standards », *MegaMathematics*, Los Alamos National Laboratory, Computer Research and Applications Group. Adresse Internet (consulté le 4 mai 2001).

CENTER FOR SUPPORT OF TEACHING AND LEARNING. 2000. « Constructing Bar Graphs », *Self-instructional Mathematics Tutorials*, Syracuse University. Adresse Internet (consulté le 6 septembre 2001).

CLARION RESEARCH. 1998. *Quality Control Procedures*. Adresse Internet (consulté le 13 septembre 2001).

CROWSON, Scott. « Hot economy stalls murder rate: Calgary second lowest in country », *The Calgary Herald*, jeudi 19 octobre 2000 (nouvelles, p. B1).

D

DELLER, Stephen. 1988. « The History of the Canadian Census », *Families*, vol. 23, nº 3, Toronto, Ontario Genealogical Society.

DOUCET, Ed, et Edvard OUTRATA. 1998. « Impact of new processing techniques on the management and organization of statistical data processing », *Journal of the United Nations Economic Commission for Europe*, ECE 5, p. 201 à 221.

DROTT, Carl M. 1994. *Dr. Drott's Random Sampler: Using the Computer as a Tool for Library Management*, College of Information Science and Technology, Drexel University. Adresse Internet (consulté le 11 juin 2001).

E

EASTON, Valerie J., et John H. MCCOLL. 1997. « Presenting Data », *STEPS Statistics Glossary Web*, version 1.1. Adresse Internet (consulté le 4 mai 2001).

ERICKSON, Bonnie H., et T.A. NOSANCHUK. 1992. *Understanding Data*, 2e éd., Toronto, University of Toronto Press.

F

FELLEGI, Ivan P. 1967. *Computer Methods for Geographical Coding and Retrieval of Data in the Dominion Bureau of Statistics*, produit nº STC3038 au catalogue de Statistique Canada, Bureau fédéral de la statistique, Division du recensement, Ottawa.

FELLEGI, Ivan P., et Simon A. GOLDBERG. 1971. *The Computer and Government Statistics*, produit nº STC2545 au catalogue de Statistique Canada, Ottawa.

FELLEGI, Ivan P., et D. HOLT. 1976. *A Systematic Approach to Automatic Edit and Imputation*, produit nº STC3432B au catalogue de Statistique Canada, Ottawa.

FERBER, Robert. 1980. *What is a Survey?* Subcommittee of the Section on Survey Research Methods, American Statistical Association (ASA), Washington, DC, ASA.

FILLIBEN, James. 1998. « Scatter Plot », *Engineering Statistics Handbook*, version précommercialisation, National Institute of Standards and Technology. Adresse Internet (consulté le 7 août 2001).

FINANCIAL FORECAST CENTRE. 1997. *What is the Correlation Coefficient?* 10 Year T-Note Forecast, Applied Reasoning Incorporated. Adresse Internet (consulté le 25 novembre 2002).

FRIEDMAN, Herbert. Introduction to Statistics.

G

GALLOWAY, Allison. 1997. *Sampling: A Workbook*. Adresse Internet (consulté le 11 juin 2001).

GARSON, David G. 2001. « Sampling », *Syllabus for PA 765: Quantitative Research in Public Administration*, North Carolina State University. Adresse Internet (consulté le 31 mai 2002).

GOWER, Allen, et Karen KELLY. 1993. *How Big Should the Sample Be?*, produit nº 10H0046XPF au catalogue de Statistique Canada, Ottawa.

GRAVETTER, Frederick J., et Larry B. WALLNAU. 1999. *Essentials of Statistics for the Behavioral Sciences*, 3 e éd., Pacific Grove, CA, Brooks/Cole Publishing Company.

H

HAUSER, Philip M. 1975. *Social Statistics in Use*, New York, Russell Sage Foundation.

HUFF, Darrell. 1954. *How to Lie with Statistics*, New York, W.W. Norton and Company.

I

INSTITUT VANIER DE LA FAMILLE. 1994. *Profil des familles canadiennes*, Ottawa, Institut Vanier de la famille.

J

JOHNSON, Richard Arnold, et Jane F. GENTLEMAN. 1989. *Améliorations apportées au SAS de l'ordinateur central en vue de faciliter l'analyse exploratoire des données*, produit nº 11F0019F nº 24 au catalogue de Statistique Canada, Direction des études analytiques, Ottawa.

JONES, James. 2001. « Estimating the Population Mean », *Lecture Notes for Math 170: Introduction to Statistics*, Richland Community College. Adresse Internet (consulté le 4 mai 2001).

— 2001. « Frequency Distributions & Graphs », *Lecture Notes for Math 170: Introduction to Statistics*. Adresse Internet (consulté le 4 mai 2001).

— 2001. « Introduction to Estimation », *Lecture Notes for Math 170: Introduction to Statistics*. Adresse Internet (consulté le 4 mai 2001).

— 2001. « Introduction to Probability », *Lecture Notes for Math 170: Introduction to Statistics*. Adresse Internet (consulté le 4 mai 2001).

— 2001. « Measures of Central Tendency », *Lecture Notes for Math 170: Introduction to Statistics*. Adresse Internet (consulté le 4 mai 2001).

— 2001. « Stats: Estimating the Mean », *Lecture Notes for Math 170: Introduction to Statistics*. Adresse Internet (consulté le 4 mai 2001).

— 2001. « Stats: Frequency Distribution and Graphs », *Lecture Notes for Math 170: Introduction to Statistics*. Adresse Internet (consulté le 4 mai 2001).

— 2001. « Stats: Measures of Central Tendency », *Lecture Notes for Math 170: Introduction to Statistics*. Adresse Internet (consulté le 4 mai 2001).

K

KAUSHAL, Ritu, et Normand J.D. LANIEL. 1993. *Computer Assisted Interviewing Data Quality Test*, produit nº 11N0022 au catalogue de Statistique Canada, Ottawa.

KENDALL, Sir Maurice G., et William R. BUCKLAND. 1982. *A Dictionary of Statistical Terms*, 4e éd. rev. et augm., London, RU, Longman Group.

KEY, James P. 1997. « Module R8-Sampling », *Research Design in Occupational Education*. Adresse Internet (consulté le 13 septembre 2001).

KIRKMAN, Tom. 2001. « Descriptive Statistics », *Tools for Science*. Adresse Internet (consulté le 11 juin 2001).

KLINKE, Sigbert. 1997. *Data Structures for Computational Statistics*, Berlin, Physica-Verlag.

KRISTULA, Dave. 2001. *The History of the Internet*. Adresse Internet (consulté le 9 septembre 2001).

KROTKI, Karol J. 1977. *Census of Computer Software at Statistics Canada*, produit nº STC1210 au catalogue de Statistique Canada, Ottawa.

L

Lane, David M. « Arithmetic Mean. » *Hyperstat Online*. 1993. (consulté le 24 novembre 2000).

— « Box plot », *Hyperstat Online*. 1993. Adresse Internet (consulté le 24 novembre 2000).

— « Linear relationship », *Hyperstat Online*. 1993. Adresse Internet (consulté le 24 novembre 2000).

— « Nominal scale », *Hyperstat Online*. 1993. Adresse Internet (consulté le 24 novembre 2000).

— « Sampling fluctuation », *Hyperstat Online*. 1993. Adresse Internet (consulté le 13 septembre 2001).

— « Scatterplots », *Hyperstat Online*. 1993. Adresse Internet (consulté le 24 novembre 2000).

— « Stem and leaf plots », *Hyperstat Online*. 1993. Adresse Internet (consulté le 24 novembre 2000).

LEINER, Barry M., et autres. 2000. *A Brief History of the Internet*, version 3.31, Internet Society. Adresse Internet (consulté le 9 septembre 2001).

LEMESHOW, Stanley, et Paul S. LEVY. 1991. *Sampling of Populations: Methods and Applications*, 2e éd., New York, John Wiley & Sons.

LEMESHOW, Stanley, et Paul S. LEVY. 1999. *Sampling of Populations: Methods and Applications*, 3e éd., New York, John Wiley & Sons.

LOHR, Sharon L. 1999. « Chapter 6: Sampling with Unequal Probabilities », *Sampling: Design and Analysis*, Pacific Grove, CA, Duxbury Press.

LOI SUR LA STATISTIQUE, L.R.C. 1985, c. S-19.

LOVEDAY, Robert. 1958. *A first course in statistics*, Cambridge, RU, Cambridge University Press.

M

MARTIN, Marjorie. 1986. *Introduction to Interviewing*, produit nº STC1595 au catalogue de Statistique Canada, Division des opérations des enquêtes, Ottawa.

MATHIEU, Richard G., et Omar KHALIL. 2001. « Data Quality in the Database Systems Course », *Data Quality*. Adresse Internet (consulté le 13 septembre 2001).

MENZIES, Heather. 1982. *Computers on the Job: Surviving Canada's Microcomputer Revolution*, Toronto, James Lorimer and Company.

MEYERS, Jeremy. 1999. *A Short History of the Computer*. Adresse Internet (consulté le 26 novembre 1999).

MURDOCK, Everett. 1994. *History, the History of Computers, and the History of Computers in Education*, California State University Long Beach. Adresse Internet (consulté le 2 novembre 1999).

N

NATIONAL STATISTICS OFFICE, REPUBLIC OF THE PHILLIPINES. 2001. *Technical Notes on the Seasonal Adjustment of CPI*. Adresse Internet (consulté le 25 janvier 2002).

NEISWANGER, William Addison. 1947. *Elementary Statistical Methods: As Applied to Business and Economic Data*, éd. rev., New York, The Macmillan Company.

NILES, Robert. 1996. « Margin of Error », *Statistics Every Writer Should Know*. Adresse Internet (consulté le 13 septembre 2001).

NOETHER, Gottfried. 1991. *Introduction to Statistics: A Fresh Approach*, New York, Springer-Verlag.

NVISION RESEARCH. 2001. *Analysis Capabilities*. Adresse Internet (consulté le 13 septembre 2001).

O

Official site of the Canadian Football League. (consulté le 26 septembre 2002).

OTTAWA CITIZEN. « Crime by the numbers », *Ottawa Citizen*, mardi 25 juillet 2000 (éditorial, p. A14).

P

PARLOR, Margaret. 1991. *Statistiques du commerce international du Canada : d'hier à demain*, produit nº 65-001 au catalogue de Statistique Canada, Ottawa.

PATON, David G. 1996. « Editing strategies and systems used by the Canadian General Social Survey: their evolution over ten years of data collection and processing », *Survey and Statistical Computing 1996*, produit nº 12F0066XPE au catalogue de Statistique Canada, Ottawa.

PLONSKY, M. 1997. « Frequency Distributions », *Psychological Statistics*. Adresse Internet (consulté le 22 janvier 2002).

POLSSON, Ken. 1994. *Chronology of Personal Computers*. Adresse Internet (consulté le 26 novembre 1999).

R

RENCKLY, Tom. 2001. « Chapter 4: Sampling Techniques and Related Statistical Concepts », *Sampling and Surveying Handbook*, Air University. Adresse Internet (consulté le 11 juin 2001).

ROUTIO, Pentti. 2001. « Sampling », *Arteology*. Adresse Internet (consulté le 11 juin 2001).

S

SANDERS, Donald A., et autres. 2001. *Statistics: A First Course*, 1ère édition canadienne, Whitby, ON, McGraw-Hill Ryerson.

SATIN, Alvin, et Wilma SHASTRY. 1983. *L'échantillonnage : un guide non mathématique*, 2e éd., produit nº 12-602F au catalogue de Statistique Canada, Division des méthodes de recensement et d'enquêtes-ménages, Ottawa.

SHAROV, Alexei. 1996. « 2.4: Stratified Sampling », *Lecture Notes for Quantitative Population Ecology*, Virginia Tech University, Department of Entomology. Adresse Internet (consulté le 11 juin 2001).

SHESKIN, Ira M. *Survey Research for Geographers*.

SHODOR EDUCATION FOUNDATION. 1997. *What? Pie Chart*. Adresse Internet (consulté le 4 mai 2001).

STARK, Philip B. 2001. « Chapter 1: Statistics », *Statistics Tools For Internet and Classroom Instruction with a Graphical User Interface (SticiGui)*, University of California, Berkeley. Adresse Internet (consulté le 4 mai 2001).

— 2001. « Chapter 2: Measures of Location and Spread », *SticiGui*. Adresse Internet (consulté le 4 mai 2001).

— 2001. « Chapter 2: Measures of Location », *SticiGui*. Adresse Internet (consulté le 4 mai 2001).

STATA.COM. 2001. *Statistical Graphs and Display*. Adresse Internet (consulté le 7 février 2001).

STATISTIQUE CANADA. 1970. *Contenu des questionnaires du recensement du Canada de 1971*, produit nº STC1623F au catalogue de Statistique Canada, Division du recensement, Ottawa.

STATISTIQUE CANADA. 1980. *Catalogue de Statistique Canada*, produit nº 11-204F au catalogue de Statistique Canada, Division de l'assistance-utilisateurs, Ottawa.

STATISTIQUE CANADA. 1980. *Comment obtenir et utiliser les statistiques : guide général de Statistique Canada*, produit nº STC2530F au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1987. *Lignes directrices pour la qualité des produits statistiques*, 2e éd., produit nº 12-539-XPF au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA . 1987. *Statistics Canada's Policy on Informing Users of Data Quality and Methodology*, produit nº STC3171 au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1991. *Pleins feux sur l'avenir*, produit nº STC2537F au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1991. *Statistiques chronologiques sur la population active*, produit nº 71-201-XPB 2000 au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1992. *Le Recensement de 1991 en bref*, produit nº 92-305F au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1993. *75 ans à compter : l'histoire de Statistique Canada*, produit nº 11-531 au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1994. *Estimations intercensitaires révisées de la population et des familles au 1er juillet, 1971-1991*, produit nº 91-537 au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1996. *Contenu du questionnaire : le recensement de la population de 1996*, produit nº 92N0064XPF au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1997. *Guide de l'Enquête sur la population active*, produit nº 71-543-GIF au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1997. *Le Recensement de 1996 en bref*, produit nº 92-352-XPF au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1998. *Case Studies for Business Schools: Preliminary Review of Existing Statistics Canada 'Case-Like' Materials*, Statistics Canada, Marketing Division. Le ministre responsable de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1998. *La statistique des entreprises : sa raison d'être : perspective du répondant*, produit nº 10F0213XPF au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1998. *Statistique Canada, lignes directrices concernant la qualité*, 3e éd., produit nº 12-539-XIF 1998 au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1998. *Les enquêtes : de A à Z* (Atelier), Le ministre responsable de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1999. *Annuaire du Canada 1999*, produit nº 11-402-XPF au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 1999. *Statistics Canada Mainframe Computers*, Since 1960, éd. rev., Ottawa.

STATISTIQUE CANADA. 2000. « Cours de base sur les enquêtes, CBE ménages », produit nº 12N0079XPF au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 2000. *Indicateurs de l'éducation au Canada : rapport du Programme d'indicateurs pancanadiens de l'éducation, 1999*, produit nº 81-582-XPF 1999 au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 2000. *L'activité humaine et l'environnement 2000*, produit nº 11-509-XPF au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA. 2002. *Annuaire du Canada 1999* (Trousse de l'enseignant). Adresse Internet (consulté le 29 mai 2002).

STATISTIQUE CANADA. 2002. « Jean Talon : 1625-1694 », *À propos de Statistique Canada*. Adresse Internet (consulté le 29 mai 2002).

STATISTIQUE CANADA. 2002. *Les diagrammes à l'ère de l'information* (Trousse de l'enseignant). Adresse Internet (consulté le 23 mars 2000).

STATISTIQUE CANADA. 2003. Le Recensement de 2001 en bref, produit nº 92-379-XIF au catalogue de Statistique Canada, Ottawa.

STATISTIQUE CANADA et FÉDÉRATION CANADIENNE DES SCIENCES SOCIALES. 1983. *Statistique historique du Canada*, 2e éd., produit nº 11-516-XPF au catalogue de Statistique Canada, Ottawa.

SUESS, Jack. 1996. *Classifying Variables by Type*, University of Maryland. Adresse Internet (consulté le 4 mai 2001).

T

TEXAS INSTRUMENTS. 1984. *Student Calculator Math: an expanded edition of TI's popular "Great International Math on Keys"* book, Texas, Texas Instruments.

TWELKER, Paul A. 2000. *Lesson 15: Discrete Random Variables*, Trinity International University. Adresse Internet (consulté le 4 mai 2001).

W

WATTIER, Mark, J. 2001. « Descriptive Statistics », *POL 660 Course Syllabus*, Murray State University. Adresse Internet (consulté le 24 novembre 2000).

WEBOPEDIA.COM. 1996. « Column Graph », *Webopedia*, INT Media Group. Adresse Internet (consulté le 4 mai 2001).

— 1996. « Raw Data », *Webopedia*. Adresse Internet (consulté le 22 janvier 2002).

WESTGARD, James O. 2000. *The Idea of Statistical Quality Control*. Adresse Internet (consulté le 13 septembre 2001).

WESTON, Harley. 1997. « A note on standard deviation », *Math Central*. Adresse Internet (consulté le 11 juin 2001).

WHITE, Stephen. 1999. *A Brief History of Computing*. Adresse Internet (consulté le 26 novembre 1999).

WILHELM, E., R. DIBBS et Wilma SHASTRY. 1983. *Definitions of Terms Used in Survey Research*, Industrie Canada, Ottawa.

WILSON, Terry C. 1980. *Researcher's Guide to Statistics: Glossary and Decision Map*, Maryland, University Press of America.

WORTON, David Albert. 1998. *Le Bureau fédéral de la statistique : les origines et l'évolution du bureau central de la statistique au Canada, 1841-1972*, produit nº 12-582-XPF au catalogue de Statistique Canada, Kingston, McGill-Queen's University Press.

WRIGHT, Jane-Marie. 1999. *MA 22 Statistics Unit: Creating a Frequency Table and Histogram*, Suffolk County Community College, Mathematics Department. Adresse Internet (consulté le 19 février 2002).

Y

YOUTSEY, Carolyn. 1998. « Line Graphs », *Carolyn's Unit on Graphing*, University of Illinois at Urbana-Champaign. Adresse Internet (consulté le 4 mai 2001).

Les statistiques : le pouvoir des données!
Glossaire

Les définitions qui suivent visent à renseigner ceux qui ont des questions concernant les statistiques, mais qui n'ont pas besoin d'une définition hautement technique.

Nos définitions représentent parfois une grande simplification de notions très complexes. Pour obtenir une définition plus technique d'un terme, vous n'avez qu'à cliquer sur la liste de dictionnairesetdéfinitionsdeStatistiqueCanada.

A

Accessibilité

L'accessibilité reflète la disponibilité de l'information détenue par l'organisme. Elle tient compte de la pertinence du format dans lequel est offerte l'information, du support de diffusion, de la disponibilité des métadonnées (texte descriptif) et du fait que l'utilisateur doit savoir que cette information existe et connaître la façon de l'obtenir. Pour les utilisateurs, le coût d'achat de cette information par rapport à la valeur qu'ils y accordent constitue aussi un aspect de cette caractéristique.

Actualité

L'actualité de l'information reflète l'intervalle entre la disponibilité de l'information et l'événement ou le phénomène décrit. Nous devons la considérer dans le contexte d'une période de temps qui confère toute sa valeur à l'information et qui permet d'y intervenir. En général, l'actualité peut influer sur la fiabilité de l'information.

Analyse fonctionnelle

Activité consistant à décomposer un problème de traitement de données en éléments fonctionnels afin de juger de la meilleure méthode de solution. Il s'agit d'un diagramme destiné à représenter la structure par âge d'une population.

Arrondissement aléatoire

Méthode par laquelle tous les chiffres d'un tableau, y compris les totaux, sont arrondis de façon aléatoire (vers le haut ou vers le bas) à un multiple de 5 ou parfois à un multiple de 10. Cette technique protège contre la divulgation directe, résiduelle ou négative des données réelles, tout en préservant au maximum l'utilité des données.

Asymétrique

Distribution asymétrique des valeurs des données. Les valeurs situées d'un côté de la courbe de distribution ont tendance à s'éloigner davantage du « milieu » que les valeurs situées de l'autre côté.

Axe des X

Ligne numérique horizontale dans le plan à coordonnées cartésiennes.

Axe des Y

Ligne numérique verticale dans le plan à coordonnées cartésiennes.

B

Baby-boomer

De façon générale, il s'agit d'une personne née après la Deuxième Guerre mondiale, entre les années 1946 et 1966.

Balance des paiements

La balance des paiements retrace l'ensemble des transactions économiques entre les résidents du Canada et les non-résidents. Elle comprend le compte courant et le compte capital et financier. **Source :** ARCHIVÉ - *Le Quotidien*-Le vendredi 28 mai 2004

Base de sondage

Liste, carte ou caractérisation théorique des unités de la population visée par une enquête; c'est le moyen par lequel on sélectionnera les répondants. Il peut s'agir par exemple d'un annuaire téléphonique ou d'un répertoire de la ville, de la liste des membres d'une association ou d'un groupement en particulier.

Base de données

Liste structurée et ordonnée de fait ou de renseignements qui est habituellement produite par ordinateur.

Biais

Dans l'estimation de la valeur d'un paramètre d'une répartition de probabilités, c'est la différence entre la valeur prévue et la valeur réelle du paramètre.

C

Caractéristique

Propriété servant à différencier les individus d'une population donnée. La différenciation entre individus peut être soit qualitative, soit quantitative.

Centiles

Valeurs qui divisent un ensemble d'observations en 100 parties égales.

Classification

Représentation d'une population donnée organisée en catégories homogènes.

Codage

Procédé servant à convertir en nombres ou en symboles l'information incluse dans un questionnaire afin de faciliter les opérations subséquentes de traitement des données. Cela demande parfois d'interpréter des réponses et de les classer dans des résultats prédéterminés.

Codage de données

Il peut arriver que des donnéesbrutes enregistrées dans un ordinateur doivent être codées. À cette fin, on étiquette chacun des éléments de données à l'aide d'un code abrégé (habituellement numérique) afin de faciliter la manipulation des données.

Coefficient de corrélation

Indice montrant dans quelle mesure deux variables fluctuent de façon interdépendante.

Coefficient de détermination

Permet de déterminer dans quelle mesure la variabilité d'une variable donnée dépend de sa relation avec une autre variable donnée. Pour calculer le coefficient, on élève au carré la valeur de la corrélationlinéaire, r. Dans un modèle linéaire par exemple, une corrélation de 0,80 (r = 0,80) signifierait que r^2, le coefficient de détermination, équivaudrait à 0,64. En conséquence, il serait possible de prédire 64 % de la variabilité des valeurs Y en se fondant sur la relation avec les valeurs X.

Coefficient de variation

Mesure de la dispersion calculée en divisant l'écart-type d'une distribution par sa moyenne. Erreur-type d'une estimation, exprimée sous forme de ratio ou de pourcentage de l'estimation.

Cohérence

Degré selon lequel les données et l'information d'un seul programme statistique sont regroupées avec d'autres données pour former un ensemble logique et complet. Des données entièrement cohérentes présentent une uniformité logique — à l'interne, au fil du temps et d'un produit ou d'un programme à l'autre. Le cas échéant, les concepts et les populations cibles utilisés ou présentés font l'objet d'une distinction logique par rapport aux concepts et aux populations cibles semblables ou aux notions et à la terminologie courantes.

Cold deck

Méthode qui repose sur un jeu fixe de valeurs, qui couvre tous les éléments de données. Ces valeurs peuvent découler de données historiques, de l'expertise de la spécialité, etc. Un questionnaire « parfait » est créé pour répondre aux besoins complets ou partiels de l'imputation.

Compte capital et financier

Le compte capital et financier porte principalement sur les transactions reliées à des instruments financiers. L'actif et le passif financiers découlant des transactions avec les non-résidents sont présentés selon trois catégories fonctionnelles : les investissements directs, les investissements de portefeuille et tous les autres types d'investissement. Ces investissements appartiennent soit à des résidents canadiens (actif du Canada), soit à des étrangers (passif du Canada). Les transactions sont dites positives si elles représentent une entrée de capital et négatives si elles représentent une sortie de capital du Canada.
Source : ARCHIVÉ - *Le Quotidien*-Le vendredi 28 mai 2004

Compte courant

Le compte courant couvre les transactions sur les biens, les services, les revenus découlant de placements, et les transferts courants. Les transactions tels les exportations et les revenus d'intérêt correspondent à des recettes alors que les importations et les versements d'intérêt correspondent à des paiements. Le solde de ces transactions détermine si le Canada enregistre un surplus ou un déficit au compte courant.
Source : ARCHIVÉ - *Le Quotidien*-Le vendredi 28 mai 2004

Confidentialité

Dans une enquête confidentielle, le caractère privé des renseignements fournis par les différents répondants est protégé, de sorte qu'il est impossible de déduire ces renseignements personnels des résultats publiés.

Conjoints de fait

Deux personnes de sexe opposé ou de même sexe qui vivent ensemble en tant que couple, sans être légalement mariées l'une à l'autre.

Connaissance des langues officielles

Capacité de converser en français ou en anglais seulement, à la fois en français et en anglais ou dans une langue autre que les langues officielles du Canada.

Contrôle de la qualité

Ensemble des opérations visant à garantir que les erreurs introduites au cours de l'enquête n'excéderont pas les niveaux déterminés.

Corrélation directe

Dans une corrélation directe, les deux variables ont tendance à suivre la même direction. Quand l'une des variables augmente, l'autre augmente aussi.

Corrélation linéaire

Détermine dans quelle mesure des points de données correspondent à une ligne droite. Quand tous les points se trouvent sur la ligne, on parle de corrélation parfaite. Quand les points sont dispersés dans tout le diagramme, il n'y pas de corrélation.

Corrélation négative

Dans une corrélation négative, les deux variables tendent à aller dans des directions opposées. À mesure qu'une variable augmente, l'autre diminue. On peut donc aussi parler de relation inverse.

Corrélation positive

Dans une corrélation positive, les deux variables tendent à aller dans la même direction. Lorsqu'une variable augmente, l'autre aussi augmente.

D

Décennal

Qui a lieu tous les 10 ans.

Désaisonnalisation

Technique statistique utilisée pour retirer des données l'effet des fluctuations saisonnières normales, afin de mieux faire ressortir les tendances sous-jacentes. Les statistiques économiques soumises à une influence saisonnière sont parfois présentées sans tenir compte de cette influence. L'effet calculé des saisons est éliminé de ces données.

Diagramme

Représentation graphique de données (p.ex. diagrammeàbandes, diagrammelinéaire, diagrammecirculaire, histogramme, pictogramme).

Diagramme à bandes

Diagramme qui compare des barres de même largeur, mais de hauteur différente, conformément aux statistiques ou aux données qu'elles représentent. Les grahiquesàbarres sont horizontales. Un diagramme à bandes présenté de façon verticale s'appelle « diagramme à colonnes ». Exempled'undiagrammeà bandes

Diagramme à colonnes

Diagramme à bandes vertical. Voir diagrammeàbandes. Exempled'undiagrammeàcolonnes

Diagramme à tiges et à feuilles

Diagramme à ramifications montrant toutes les combinaisons ou tous les résultats possibles.

Diagramme circulaire

Diagramme circulaire permettant d'élaborer un concept visuel d'un ensemble donné (100 % = 360 degrés). Le diagramme est divisé en secteurs, chacun correspondant à une catégorie de la variable présentée (p.ex., chaque groupe d'âge). Chaque secteur est proportionnel au pourcentage de la catégorie correspondante. Exempled'undiagrammecirculaire

Voir diagrammecirculaire.

Diagramme linéaire

Diagramme où les points successifs représentant les valeurs d'une variable pour certaines valeurs de la variable dépendante sont reliés par des lignes droites (p.ex., le taux de chômage chez les jeunes au cours des 10 dernières années). Exempled'undiagrammelinéaire

Diagramme par points

Diagramme à deux dimensions représentant deux variables sous la forme d'une série de points. Principalement utilisé pour montrer la corrélation entre les deux variables. Exempled'undiagrammedepoints

Dispersion

Décrit dans quelle mesure les observations divergent autour de la tendancecentrale.

Distribution de fréquences

diagramme ou tableau montrant avec quelle fréquence chaque valeur ou chaque série de valeurs d'une variable apparaît dans un ensembledededonnées.

Distribution de fréquences groupées

Relation entre les valeurs d'une caractéristique et leurs fréquences, lorsque ces valeurs sont groupées en intervallesdeclasse.

Distribution de fréquences non groupées

Distributiondefréquences de données numériques où les donnéesbrutes ne sont pas groupées.

Distribution normale

Souvent appelée simplement « courbe en cloche » ou « courbe en forme de cloche ». La plupart des résultats de ce diagramme sont groupés au centre. La moyenne, la médiane et le mode sont égaux et les résultats situés à chaque extrémité de la distribution sont moins fréquents. Par exemple, dans une courbe représentant les résultats d'un test d'intelligence, la plupart des personnes se trouvent au centre ou autour de l'intervalle d'intelligence « moyen », tandis que le nombre de personnes diminue des deux côtés à mesure que les résultats s'éloignent de la moyenne, d'où la forme et le nom de la courbe.

Dollars constants

Dollars d'une année de référence précise, indexés en fonction de l'inflation ou de la déflation, de façon à montrer les changements du pouvoir d'achat du dollar (voir aussi « Dollarscourants »). L'année de référence doit toujours être indiquée.

Dollars courants

Dollars qui expriment le coût d'articles en fonction de l'année où ceux-ci sont achetés. Il est à noter que les « dollars courants » sont aussi nommés « dollars de l'année budgétaire », par opposition aux « dollarsconstants ».

Données

Faits ou chiffres à partir desquelles on tire des conclusions.

Données brutes

Information qui n'a pas encore été organisée, mise en forme ou analysée.

Données catégoriques

Données qui peuvent être groupées en catégories spécifiques (aussi nommées « variables qualitatives »). Les variables catégoriques peuvent être associées à des catégories qui suivent un ordre naturel (variablesordinales) ou qui ne suivent pas d'ordre naturel (variablesnominales). Par exemple, la variable « hauteur » est ordinale parce qu'elle contient les catégories « petit », « moyen » et « grand » qui suivent naturellement un ordre ascendant. Au contraire, des variables comme « sexe » et « couleur des cheveux », qui ne contiennent pas de catégories naturellement ordonnées, sont des exemples de variables nominales.

Données élémentaires (éléments d'information)

Le plus petit élément d'information qu'on puisse obtenir par une enquête ou un recensement.

E

Écart interquartile

Voir Intervalleinterquartile.

Écart semi-interquartile

Voir Intervallesemi-interquartile.

Écart-type

Racine carrée de la variance, l'écart-type mesure l'étalement ou la dispersion par rapport à la moyenne d'un ensemble de données. C'est la mesure de l'étalement la plus utilisée.

Échantillon non aléatoire

Voir Échantillonnagenonprobabiliste.

Échantillonnage à participation volontaire

Échantillons formés de personnes qui ont offert bénévolement leurs services en sachant que le processus serait long ou exigeant, et parfois même désagréable. On a souvent constaté que ces bénévoles ont une attitude favorable ou tout au moins neutre, alors que la population en général a tendance à manifester des attitudes plus diverses au sujet des questions concernées.

Échantillonnage à plusieurs degrés

Activité par laquelle on prélève un échantillon en deux ou trois étapes. Il y a dans ce cas une hiérarchie de types d'unités où chaque unité de premier degré pourrait se décomposer en unités de deuxième degré, et ainsi de suite.

Échantillonnage aléatoire Voir

Échantillonnageprobabiliste.

Échantillonnage aléatoire simple

Échantillonnageprobabiliste de base où chaque échantillon a des probabilités égales d'être tirée.

Échantillonnage à l'aveuglette

On procède à un tel échantillonnage au hasard par commodité ou disponibilité.

Échantillonnage au jugé (par choix raisonné)

Activité d'échantillonnage reposant sur l'hypothèse selon laquelle la compétence et le jugement personnels peuvent permettre une sélection d'unités typiques ou représentatives de la population cible.

Échantillonnage non probabiliste

Échantillon sélectionné par une méthode non probabiliste. Par exemple, un schéma dans lequel les unités sont choisies dans un but précis, donnerait un échantillon non aléatoire.

Échantillonnage par participation volontaire

On forme des échantillons avec des gens qui ont offert leurs services en sachant que l'exercice serait long ou exigeant et deviendrait peut-être même désagréable. On a souvent constaté que ces bénévoles avaient des attitudes sinon favorables, du moins neutres, alors que la population en général avait tendance à varier plus largement dans ses attitudes à l'égard des sujets d'intérêt.

Échantillonnage par quotas

Activité où le nombre de répondants appartenant à une catégorie (parmi plusieurs) est établi d'avance et où la sélection finale des répondants est confiée à l'intervieweur qui échantillonne jusqu'à ce que toutes les unités de chaque catégorie soient réunies.

Échantillonnage probabiliste

Méthode d'échantillonnage où chaque membre de la population a une chance d'être sélectionné. On le nomme aussi « échantillonnage aléatoire » en raison de la manière aléatoire dont les personnes sont choisies afin de garantir une représentation sans biais de l'ensemble de la population.

Échantillonnage stratifié

Procédé d'échantillonnage où la population se répartit en sous-groupes homogènes ou en strates et où le tirage d'échantillons est indépendant dans chaque strate.

Échantillonnage systématique

Sélection d'unités dans une liste à l'aide d'un intervalle de sélection (k), de façon à ce que chaque élément de la liste touché par l'intervalle k soit inclus dans l'échantillon, l'origine ayant été choisie au hasard entre 1 et k. Si par exemple k était égal à 6 et si l'origine choisie au hasard était 2, l'échantillon serait 2, 8, 14, 20, etc.

Échelle

Ligne graduée divisée en valeurs successives, pouvant être diagramme, descriptive ou numérique, utilisée pour rendre compte d'évaluations. Dans les diagrammes, l'échelle est la subdivision de chaque axe. L'échelle peut être numérique ou catégorique. (Voir aussi : Lespartiesd'undiagramme.)

Effectif

Voir fréquence.

Enquête

Collecte de données sur des caractéristiques d'intérêt d'une partie ou de la totalité des unités d'une population à l'aide de concepts, de méthodes et de procédures bien définis. Elle est suivie d'un exercice de rassemblement permettant de présenter les données recueillies sous une forme récapitulative utile.

Enquête-échantillon (ou enquête par sondage)

Collecte d'information auprès d'une partie seulement de la population.

Enquêtes polyvalentes

Enquêtes dont les objectifs prévoient une mesure d'un grand nombre de caractéristiques. Par exemple, une enquête sur les dépenses agricoles pourrait viser à établir plus que les frais totaux d'exploitation agricole et vouloir en particulier découvrir ce qu'il en coûte au total pour les instruments aratoires, les salaires, les emprunts, les semences, l'alimentation des animaux, etc.

Enregistrement magnétique

Enregistrement qui permet la lecture et l'écriture. Cette méthode est utilisée lorsqu'il importe de garantir la sécurité des données. L'application la plus répandue de ce type de saisie est la lecture du NIP des cartes bancaires.

Ensemble de données

Tout groupe de données ayant un thème commun ou des caractéristiques semblables.

Entrée de données

Procédé (numérisation, inscription sur papier, bandes magnétiques, fiches, etc.) utilisé pour relever la saisie des données.

Équation de régression

Équation qui permet de prévoir une variable inconnue en utilisant la valeur donnée d'une ou de plusieurs autres variables. Par exemple, l'équation Y = a + b X donne la valeur estimative de Y, si la valeur de X est connue. (Voir aussi Régression et Modèlederégression).

Erreur aléatoire

Erreurs imprévisibles dans une estimation. Ces erreurs ont tendance à s'annuler dans un grand échantillon, contrairement aux erreurs systématiques qui vont toutes dans la même direction.

Erreur d'échantillonnage

Erreur qui se produit lorsque les données sont recueillies auprès d'une partie seulement de la population. Elle est habituellement mesurable par les données de l'échantillon en cas d'échantillonnageprobabiliste.

Erreur de non-réponse

Erreur causée par l'absence d'interview ou de non-réponse à une question d'un questionnaire.

Erreur non due à l'échantillonnage

Erreur causée par des facteurs autres que l'échantillonnage (des erreurs de couverture, des erreursderéponse, des erreursdenon-réponse, des questionnaires erronés, ainsi que des erreurs de consignation par les intervieweurs, de traitement, etc.).

Erreur de réponse

Différence entre la vraie réponse à une question et la réponse donnée par le répondant; elle peut être imputable au répondant, à l'intervieweur, au questionnaire, au procédé d'enquête ou à l'interaction entre le répondant et l'intervieweur.

Erreur de traitement

Erreur qui se glisse dans un exercice de transformation des données de questionnaires, de listes de contrôle, etc. en des ensembles de totalisations et d'estimations.

Estimation

Fait de tirer des conclusions plus vastes d'un échantillon afin de prédire certaines caractéristiques ou tendances de l'ensemble de la population.

État matrimonial

Représente l'état de personnes mariées devant la loi (et non séparées), de personnes en union libre, de personnes séparées (mais toujours mariées devant la loi), de personnes divorcées, de veufs ou de veuves et de célibataires (personnes jamais mariées).

Étendue

Voir Fourchette.

Exactitude

Mesure dans laquelle les résultats d'un calcul ou de la lecture d'un instrument traduisent les quantités réelles qui font l'objet de ce calcul ou de cette lecture et sont exempts d'erreurs. (Voir aussi Précision)

Exclusif

Lorsqu'un événement exclut automatiquement la possibilité qu'un second événement se produise au même moment, les résultats sont exclusifs. Par exemple, rond et carré sont des termes exclusifs, puisqu'un objet ne peut pas avoir ces deux formes au même moment.

Exhaustif

Quand un ensemble d'événements inclut toutes les occurrences possibles d'un ensemble de références, les résultats sont exhaustifs. Par exemple, la liste de groupes d'âge 0-19, 20-34, 35-59 et 60 ans et plus est exhaustive, parce qu'elle couvre l'éventail des âges que peuvent avoir les membres de la population.

F

Fluctuation d'échantillonnage

Fait qu'une statistique prend des valeurs différentes pour des échantillons différents. Elle indique donc dans quelle mesure la valeur de la statistique fluctue d'un échantillon à l'autre.

Formule

Équation ou règle mathématique.

Fourchette

Tout l'intervalle de variation numérique des résultats. La fourchette exclusive est la différence entre la valeur la plus élevée et la valeur la moins élevée d'un ensemblededonnées, et la fourchette inclusive, la différence entre la limite supérieure réelle de l'intervalle le plus long et la limite inférieure réelle de l'intervalle le plus court.

Fréquence

Nombre de fois qu'un événement se produit ou qu'un élément se présente dans un ensemblededonnées.

Fréquence cumulée

Détermine le nombre d'observations sous-jacent à une valeur particulière. On la calcule en ajoutant chaque fréquence à la somme de ses prédécesseurs dans un diagrammearborescent et un tableau de distribution de fréquences.

Fréquence en pourcentage

Fréquence de chaque valeur ou de chaque intervalle de classe, exprimée sous forme de pourcentage du nombre total d'observations. Obtenue en multipliant chacune des valeurs de la fréquencerelative par 100.

Fréquence relative

Fréquence d'une valeur ou d'un intervalledeclasse, exprimée en tant que proportion d'un tout et observée dans un ensemble de données pour une variable particulière. On la calcule en divisant la fréquence par le nombre d'observations.

Fréquentation scolaire

Fréquentation à temps plein ou à temps partiel (de jour ou de soir) d'une école, d'un collège ou d'une université pendant les huit mois compris entre les mois de septembre et de mai (jour du recensement). Cette fréquentation se mesure uniquement par rapport aux études menant à un certificat, à un diplôme ou à un grade.

G

Graphique (diagramme) arborescent

Voir diagrammesàtigesetàfeuilles. **Grappe**

Ensemble d'unités d'une population qu'on constitue à l'aide de critères bien définis. Il peut s'agir d'un groupe qui existe dans la population (îlot urbain, hôpital, etc.) ou d'un groupe théorique (aires d'un quadrillage de carte, etc.).

Groupe de discussion

Technique d'interview où l'on interroge les répondants en groupe. Cette technique permet de les amener à parler librement, à exposer leurs idées et à faire spontanément part de leurs attitudes et de leurs sentiments sur une question. On s'en sert souvent pour explorer un éventail de questions avec les réactions des répondants en vue de bien orienter la conception d'un questionnaire.

H

Histogramme

Diagramme formé d'une suite de colonnes avec un intervalledeclasse en abscisse et une courbe de fréquences en ordonnée. Exemple

Hot deck

La méthode du hot deck recourt à d'autres enregistrements (donneurs) pour répondre à la question (ou à l'ensemble de questions) qui doit faire l'objet d'une imputation. Le donneur peut être choisi au hasard parmi un groupe de donneurs affichant le même jeu de caractéristiques prédéterminées. Par exemple, dans le cas où un questionnaire serait retourné parce qu'il manquait le revenu dans un enregistrement, nous pourrions regrouper toutes les unités déclarantes de la même province, de la même profession et du même nombre d'années d'expérience. On pourrait ensuite dresser une liste des donneurs qui répondent à ces critères, puis en choisir un au hasard. Le revenu déclaré par cette personne tiendrait lieu de revenu pour la réponse manquante ou invalide.

I

Imputation

Remplacement de données manquantes ou non valides par des données acceptées. Normalement effectuée conformément à des règles de décision préétablies. Elle est souvent combinée à la vérificationdesdonnées.

Indice

Désignation mathématique ou nombre servant à exprimer le niveau de prix, le volume du commerce, le montant relatif, etc. d'une période donnée, comparativement à celui d'une période de base, par exemple, l'indice des prix à la consommation.

Industrie

Groupe de fabricants ou de fournisseurs de services qui ont été rassemblés en fonction de l'homogénéité de leurs produits ou services.

Information

Données qui ont été enregistrées, classées, organisées, raccordées ou interprétées dans un cadre quelconque pour que la signification s'en dégage.

Internet

Réseau informatique mondial reliant des ordinateurs par satellite et par téléphone qui permet aux utilisateurs de se connecter à des réseaux comme le courriel et le W3 (World Wide Web).

Intervalle

Ensemble de chiffres se situant entre deux chiffres limites et pouvant comprendre l'un ou l'autre de ces chiffres; ainsi, l'intervalle [1,5;3] comprend tous les chiffres supérieurs ou égaux à 1,5 et inférieurs à 3. Il convient de noter que le chiffre 3 se trouve exclu de cet intervalle.

Intervalle de confiance

Estimation utilisant une série de valeurs (un intervalle) pour prédire la valeur probable d'un paramètre inconnu, accompagnée d'un niveau de confiance spécifique, ou probabilité que l'estimation sera exacte (c.-à-d. que l'intervalle contiendra effectivement la valeur vraie du paramètre).

Intervalles de classe

Si une variable comporte un grand nombre de valeurs, il est plus facile, pour présenter les données, en groupant les valeurs en intervalles de classe (par exemple, groupes d'âge tels que 0-4, 5-9, 10-14, 15-19, etc. pour rendre compte de l'âge de la population), que de présenter toutes les valeurs séparément. Cela permet de distinguer plus facilement les tendances des données.

Intervalle interquartile

Différence entre les quartiles supérieur et inférieur (Q3-Q1) d'un ensembledededonnées. Cet intervalle sert à mesurer l'étalement des données : il s'étend sur 50 % d'un ensemble de données et élimine l'influence des valeursaberrantes (suppression des quarts supérieur et inférieur d'un ensemble de données).

Intervalle semi-interquartile

Estimé à la moitié de la différence entre le 75e (Q3) et le 25e centile (Q1). La formule est : (Q3-Q1)÷2. Comme la moitié des valeurs se situent entre Q3 et Q1, l'intervalle semi-interquartile équivaut à la moitié de la distance nécessaire pour couvrir la moitié des valeurs.

Intranet

Réseau informatique interne privé, l'intranet allie la commodité d'Internet au rendement et à la sécurité d'un système maison. Un intranet peut être relié à Internet au moyen d'une passerelle protégée, munie de pare-feux pour empêcher tout accès extérieur non autorisé aux systèmes internes.

L

Langue maternelle

Première langue apprise à la maison durant l'enfance et encore comprise par la personne au moment du recensement.

Langue parlée à la maison

Langue parlée le plus souvent à la maison par le répondant au moment du recensement.

Lecteur de reconnaissance

Connu aussi sous le nom « lecteur de code à barres », cet appareil peut reconnaître des lettres et des chiffres. Ces lecteurs balaient chaque ligne et les convertissent dans le programme.

Logiciel

Il y a deux catégories de logiciels : les logiciels de base (qui contrôlent le fonctionnement d'un ordinateur en Windows, DOS, etc.) et les logiciels d'application (Word, EXCEL, MS ACCESS).

M

Manipulation des données

Activité par laquelle on met les données en forme définitive (diagramme, moyennes arithmétiques, données agrégées, etc.).

Marge d'erreur

Nombre relatif qui peut être exprimé sous forme de pourcentage et qui est calculé à l'aide de l'erreur d'échantillonnage d'une estimation. On l'utilise pour établir l'intervalledeconfiance de cette estimation.

Médiane

Valeur de l'élément central lorsque les données sont ordonnées de la plus basse à la plus haute; mesure de la tendancecentrale. Si le nombre d'observations est pair, la médiane est la moyenne des deux observations centrales. Dans le cas des donnéesbrutes, la médiane correspond au point par rapport auquel la moitié exacte des données lui est supérieure et l'autre, inférieure.

Méthode de l'estimateur

Méthode qui recourt à une information tirée d'une autre question ou d'une autre réponse d'un enregistrement qui fera l'objet d'une imputation (du cycle actuel ou d'un cycle antérieur) et, au moyen d'opérations mathématiques, produit une valeur plausible destinée au champ vierge ou erroné.

Méthode de la substitution

Méthode qui repose sur la disponibilité de données comparables. Les données imputées peuvent être extraites de l'enregistrement du répondant obtenu lors d'un cycle antérieur de l'enquête ou d'un autre fichier source (p.ex., fichiers administratifs ou autres fichiers d'enquête du même répondant). Cette méthode est souvent difficile à réaliser parce que, dans de nombreux cas, il n'existe aucune autre information au sujet de ce répondant, à part l'information de l'enquête en cours.

Méthodologie

Ensemble de méthodes et de techniques de recherche appliquées à un domaine d'études particulier. À Statistique Canada, le terme « méthodologie » signifie « techniques d'enquête ».

Mode

Observation la plus fréquente dans un ensembledededonnées; il s'agit d'une mesure de la tendancecentrale.

Modèle de régression

Modèle statistique utilisé pour décrire la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Ces modèles présentent des formes et des degrés de complexité très divers. (Voir aussi Régression et Équationderégression).

Moyenne

Mesure la plus courante de la tendancecentrale, la moyenne est la moyenne arithmétique d'un ensemble de nombres.

Municipalité dissoute

Il est possible que les limites et les noms des subdivisions de recensement fassent l'objet de modifications d'un recensement à l'autre par suite d'annexions, de dissolutions et d'incorporations. Ces changements peuvent résulter en la « dissolution » de différentes subdivisions de recensement. Une subdivision de recensement dissoute est une communauté qui existait le 1er janvier 1996, mais qui n'existait plus le 1er janvier 2001, soit la date de référence géographique du recensement de 2001. Le concept de « Subdivision de recensement - Recensement précédent » a été créé pour permettre la totalisation des données du recensement actuel selon les subdivisions de recensement du recensement précédent (1996). On procède au couplage des îlots du recensement de 2001 avec les subdivisions de recensement du recensement de 1996 selon un « ajustement optimal », afin d'être en mesure de totaliser les données du recensement actuel selon les communautés du recensement précédent.

N

Non-réponse

Situation dans laquelle on ne dispose pas de données sur certaines unités de l'échantillon pour une raison quelconque. Il se peut, par exemple, que le répondant ne soit pas disponible ou qu'il refuse de répondre ou de prendre part à l'interview.

O

Observations

Données recueillies pour une variable quelconque.

Ogive

Courbe d'un diagramme de distribution de fréquences. Il est à noter que les courbes de distribution n'ont pas toutes la forme d'une ogive. Il est donc préférable de limiter ce terme à la distributionnormale ou quasi normale.

Ordinateur central

Ordinateur offrant des possibilités et des ressources étendues et auquel d'autres ordinateurs peuvent être connectés de manière à partager les ressources.

Ordinateur de milieu de gamme

Éventail très vaste d'ordinateurs allant des ordinateurspersonnels haut de gamme aux ordinateurscentraux. Antérieurement appelés « mini-ordinateurs » et pourvus de terminaux non intelligents connectés à des systèmes centralisés, la plupart des ordinateurs de milieu de gamme sont utilisés aujourd'hui comme serveurs dans un environnement client-serveur.

Ordinateur personnel

Micro-ordinateur monoposte à usages multiples, conçu pour être utilisé par une seule personne à la fois.

Origine ethnique

Groupes ethniques ou culturels auxquels appartient l'ascendance d'un répondant; par origine ethnique ou culturelle, on entend les « racines » ethniques ou l'ascendance d'une population, terme à ne pas confondre avec les termes citoyenneté et nationalité.

P

Paramètre

Il s'agit de valeurs inconnues quantitatives (p.ex., recettes totales, recettes moyennes, rendement total, nombre de chômeurs) qui concernent toute la population ou certains domaines d'intérêt pour le chercheur.

Pertinence

La pertinence des données ou de l'information statistique constitue une évaluation quantitative de leur valeur inhérente. Cette valeur dépend du degré selon lequel les données et l'information conviennent aux buts de leur production et à l'intérêt qu'y portent les utilisateurs. La valeur dépend aussi du mandat de l'organisme, des exigences imposées par la loi et du coût de substitution de la production de ces données ou de cette information.

Peuples autoch-tones

Indiens, Métis ou Inuit (Esquimaux) d'Amérique du Nord.

Pictogramme

Diagramme de représentation imagée de statistiques (on peut utiliser, par exemple, un billet de 1 dollar en grandeur croissante pour représenter l'augmentation du pouvoir d'achat dans le temps). Exemple

Plan d'échantillonnage

Ensemble de spécifications qui détaillent la population, la basedesondage, les unitésd'enquête, la taille de l'échantillon et les méthodes d'échantillonnage et d'estimation.

Polygone de fréquences

Diagramme que l'on forme en joignant en leurs points milieux les courbes en colonnes d'un histogramme (Voir aussi Histogramme)

Population

Ensemble des unités visées par les résultats de l'enquête (les unités en question sont des personnes, des animaux, des objets, des entreprises, des voyages, etc...)

Population active

Population de 15 ans et plus — à l'exception des pensionnaires d'établissement — qui est présente sur le marché du travail la semaine où tombe le 15e du mois précédant le jour du recensement. On classe les répondants en tant que personnes occupées (ayant un emploi) ou chômeurs. Le reste de la population en âge de travailler est alors considéré comme inactif.

Possibilité d'interprétation

La possibilité d'interprétation reflète la facilité avec laquelle l'utilisateur peut comprendre, utiliser et analyser les données ou l'information. La justesse des définitions des concepts, des populations cibles, des variables et de la terminologie sous-jacente aux données, ainsi que l'information sur toute limitation des données, détermine en grande partie la possibilité d'interprétation.

Pourcentage cumulé

Permet de calculer la fréquencecumulée en divisant ce pourcentage par le nombre d'observations et en le multipliant par 100. La dernière valeur est toujours égale à 100 %. Cela facilite la comparaison des données.

Précision

La précision est une mesure de la similarité. Le même sondage exécuté plus d'une fois devrait donner les mêmes résultats ou des résultats similaires. Plus les résultats de chaque répétition du sondage se ressemblent, plus ils sont précis.

Programmation

Production d'un jeu d'instructions pour qu'un ordinateur exécute une tâche en particulier.

Pyramide des âges

Diagramme qui vise à représenter la structure par selon l'âge et le sexe d'une population. Il est formé de deux histogrammes horizontaux qui sont joints. Exemple d'un pyramide des âges

Q

Qualité des données

Niveau ou degré de confiance que les données et l'information statistique sont « prêtes à l'emploi ». Les aspects auxquels Statistique Canada doit prêter attention sur le plan de la qualité ou de la capacité d'utilisation sont la pertinence, l'exactitude, l'actualité, l'accessibilité, la possibilité d'interprétation et la cohérence.

Quartiles

Pour déterminer l'intervalle interquartile, on divise l'ensemble de données en quatre parties égales. Chaque valeur marquant un palier est appelée quartile (premier, deuxième, etc.). Le deuxième quartile est aussi appelé la médiane.

Questionnaire

Série de questions conçues pour l'obtention de renseignements sur un ou plusieurs sujets auprès d'un répondant.

R

Rang-centile

Dans une distribution, proportion de valeurs à laquelle une valeur spécifique est supérieure ou égale. Par exemple, si une personne obtient une note de 95 % à un test de mathématiques et que cette note est supérieure ou égale aux notes obtenues par 88 % des élèves, la personne se trouve dans le 88e centile.

Ratio

Rapport proportionnel entre deux quantités ou deux nombres différents ou, en mathématique, quotient de deux nombres ou de deux expressions obtenu en divisant l'un des éléments par l'autre.

Recensement

Collecte de données sur toutes les unités d'une population, ce que l'on appelle parfois « enquête sur échantillon ». (Le terme « Recensement » [avec la majuscule initiale] désigne habituellement le Recensement de la population [à l'échelle nationale].)

Recensement décennal (Canada)

Les recensements ont lieu au début de chaque décennie, lors les années se terminant par le chiffre 1. Statistique Canada, par exemple, a effectué des recensements à tous les dix ans, de 1871 à 2001. (Voir aussi Recensement quinquennal)

Recensement quinquennal (Canada)

Recensement qui se tient à la mi-décennie dans les années se terminant par le chiffre 6. Par exemple, Statistique Canada a tenu des recensements quinquennaux en 1976, 1986 et 1996. (Voir aussi Recensement décennal)

Région métropolitaine

Statistique Canada a créé des groupements de municipalités, ou subdivisions de recensement, qui englobent le secteur placé sous l'influence d'un grand centre urbain. Le regroupement de ces municipalités étroitement interreliées parce que les habitants travaillent dans une municipalité et résident dans une autre, est régi par des lignes directrices précises. Les unités géographiques qui en résultent sont appelées « régions métropolitaines de recensement ».

Régression

Méthode statistique par laquelle on essaie de prévoir la valeur d'une caractéristique en étudiant sa relation avec une ou plusieurs autres caractéristiques. Cette relation s'exprime au moyen d'une équation de régression. (Voir aussi Modèle de régression)

Réseau étendu (RE)

Réseau de communication couvrant un vaste secteur géographique tel qu'une province ou un pays. Il diffère du réseau local (RL) qui se limite à un bâtiment ou à un ensemble de bâtiments, ainsi que du réseau de la région métropolitaine, qui couvre généralement une ville ou une banlieue.

Réseau local (RL)

Réseau de communication desservant les utilisateurs d'un secteur géographique limité. Il se compose de serveurs, de postes de travail, d'un système d'exploitation de réseau et d'un lien de télécommunication.

Révision

Voir révisiondedonnées. (Voir aussi imputation)

Révision de données

Procédé destiné à garantir que les données d'enquête sont exactes, complètes et uniformes.

S

S

Symbole mathématique de l'écart-type.

S²

Symbole mathématique de la variance.

Saisie des données

Procédé consistant à enregistrer des réponses sur un support lisible par machine.

Saisie interactive

Procédé souvent désignée saisie intelligente. En général, les données saisies font l'objet d'une vérification avant leur imputation. Toutefois, cette méthode combine la saisie et la vérification en une seule fonction. Bien que la saisie interactive soit plus lente, elle est très efficace lorsqu'il y a une étroite interdépendance entre les questions. Ce processus nécessite une connaissance des procédures de vérification, car il importe de corriger les erreurs sur le champ. La saisie interactive réduit aussi le nombre de documents traités, car les vérifications sont faites directement par ordinateur.

Saisie par lots

Procédé qui constitue l'une des plus anciennes méthodes de saisie. Elle recourt à un clavier pour entrer les données. Ce processus convient parfaitement à l'entrée d'une grande quantité de données lorsque la rapidité est une nécessité. Aucune procédure de vérification n'a lieu, mais nous devons placer une bonne part de la confiance dans le programme de vérification. En outre, nous devons mettre en oeuvre des vérifications de la validité et des étendues pour veiller à la qualité de la saisie. Cela ne signifie pas que nous modifions les données, mais si un champ numérique contient plutôt des lettres, l'erreur sera détectée. Cette approche peut être avantageuse dans le cadre de grandes enquêtes comportant de nombreuses questions et vérifications.

Secteur de dénombrement (SD)

Région parcourue par un intervieweur dans le cadre du Recensement de la population du Canada. Les SD sont des unités bien définies et facilement reconnaissables sur les cartes et sur le terrain, mais qui sont propres à un recensement.

Σ
(sigma)

Symbole mathématique de sommation de valeurs.

Sous-produit administratif

Données disponibles à partir de l'information enregistrée dans les dossiers administratifs, les demandes, les rapports, etc.

Statistique

Fonction produisant un résultat numérique dont la valeur peut varier selon les résultats d'une expérience (p.ex., avec des échantillons différents). Exemple : la moyenne ou la médiane d'un échantillon, etc.

Statistiques

Type d'information obtenue en soumettant les valeurs numériques à des opérations mathématiques.

Statistiques, études des

Domaine d'investigation ayant pour objet la collecte et l'agencement de données, de faits ou de valeurs numériques se rapportant aux affaires humaines ou aux phénomènes naturels.

Statistiques de l'état civil

Statistiques sur les naissances, les décès, les mariages, la santé et la maladie.

Statistique inférentielle

Méthodes statistiques utilisées pour inférer les valeurs d'une population à partir des valeurs obtenues pour l'échantillon.

Les statistiques : le pouvoir des données!

Une ressource de Statistique Canada qui a été élaborée de façon à pouvoir tirer pleinement parti des statistiques.

Stockage de données

1) Capacité d'un ordinateur à emmagasiner de l'information.
2) Il s'agit également d'une fonction informatique de mise en mémoire de l'information sur un support (p.ex., bande magnétique, disquette, CD-ROM, etc.).

T

Tableau de pointage

Sert à enregistrer les données d'une expérience, à compter les occurrences d'un événement et à élaborer les tableaux de distributiondefréquences.

Tableau statistique

Présentation de données statistiques regroupant les valeurs d'une caractéristique ainsi que le nombre de fois que chaque valeur apparaît dans l'ensemblede données.

Taux d'activité

Pourcentage de la population âgée de 15 ans et plus.qui fait partie de la populationactive. Le taux d'activité d'un groupe en particulier (p.ex., celui des femmes de 25 ans et plus) s'exprime en pourcentage de la population totale du groupe en question.

Tendance centrale

Mesure de l'emplacement où se trouve le milieu ou le centre d'une distribution. La tendance centrale peut se rapporter à des mesures très diverses comme la moyenne, la médiane et le mode. La moyenne est la mesure de la tendance centrale la plus fréquemment utilisée.

(Diagrammes à) tiges et à feuilles

Méthode semi-graphique utilisée pour représenter des données numériques, où le premier chiffre (le plus à gauche) de chaque valeur correspond à la tige, tandis que le reste des chiffres qui composent le nombre sont les feuilles. Un diagramme arborescent montre toutes les valeurs des données d'un ensemble-échantillon.

Traitement des données

Procédé consistant à convertir des donnéesbrutes sous une forme lisible par machine, puis à les trier, à les vérifier, à les manipuler et à les présenter de manière à créer de l'information.

U

Unité centrale de traitement

L'unité centrale de traitement (UCT) est le cœur d'un système informatique. Elle est assez petite pour tenir dans une main, mais peut contenir des millions de circuits logiques.

Unité d'entrée

Organe ou dispositif (bande, carte, clavier, disquette, CD-ROM, photostyle, lecteur optique, caméra numérique, etc.) à partir duquel des données peuvent être introduites dans un ordinateur.

Unités d'enquête

Aux fins du tirage d'un échantillon, la population devrait être divisible sans recoupement en un nombre fini d'unités distinctes et identifiables de sorte que chaque membre de la population ne corresponde qu'à une unité d'échantillonnage.

V

Valeurs aberrantes

Dans un ensemble de données, ce sont les valeurs si éloignées des autres valeurs d'une distribution que leur présence ne saurait être attribuée aux seules probabilités.

Variable

Caractéristique susceptible de prendre plusieurs des valeurs d'un ensemble auquel une mesure numérique peut être appliquée (p.ex., revenu, âge, poids).

Variable continue

Variablenumérique qui peut prendre un nombre infini de valeurs réelles. L'âge, la distance et la température par exemple sont considérés comme des variables continues, parce qu'une personne peut parcourir 3 642 531 km à pied.

Variable discrète

Variablenumérique qui prend uniquement un nombre limité de valeurs réelles (par ex., X peut seulement être égal à 1, 3, 5 et 1000).

Variable groupée

Ensemble de données qui a été groupé ou classé en fonction de caractéristiques qualitatives ou quantitatives communes. Exempled'unevariablegroupée

Variable nominale

Type de variable catégorique qui décrit un nom, une étiquette ou une catégorie sans ordre naturel. Par exemple, la liste des différents types de matières scolaires ne présente pas d'ordre naturel : l' « histoire » ne suit pas nécessairement la « biologie ». Ces matières peuvent être placées dans n'importe quel ordre.

Variable non groupée

Ensemble de données qui n'ont pas été groupées, ni classées; il s'agit plutôt d'une liste de valeurs individuelles observées (p.ex., années d'âge).

Variable numérique

Variable quantitative qui décrit une valeur mesurée de façon numérique (par ex., âge ou nombre de personnes dans un ménage). Ces variables peuvent être continues ou discrètes.

Variable ordinale

Type de variable catégorique : une variable ordinale présente un ordre naturel de ses valeurs possibles, mais les distances entre les valeurs ne sont pas définies. Habituellement, les variables ordinales ont des échelles catégoriques. Par exemple, lorsqu'on demande à des personnes de choisir entre « excellent », « bon », « moyen » et « médiocre » pour évaluer quelque chose, la réponse indique seulement une catégorie, mais il existe un ordre naturel de ces catégories.

Variance

Mesure de l'étalement équivalant au carré moyen de l'écart de chaque nombre par rapport à la moyenne d'un ensemble de données. Le terme variance est aussi lié à la variationd'échantillonnage.

Variation d'échantillonnage

Variation qu'accusent des échantillons d'une même population qui sont différents, mais d'une même taille.

Vérification

Activité par laquelle on vérifie les valeurs d'une variable dans tous les enregistrements ou des combinaisons préétablies de zones de tous les enregistrements pour ainsi s'assurer que les valeurs indiquées sont valables (Voir aussi Imputation)

Vérification de la cohérence

Activité par laquelle on compare les diverses réponses d'un enregistrement pour en assurer la cohérence l'une par rapport à l'autre. Par exemple, si une personne déclare faire partie du groupe d'âge de 0 à 14 ans et est à la retraite, il existe alors un problème de cohérence entre les deux réponses. La vérification inter-champ constitue une autre forme de vérification de la cohérence. Elle vise à vérifier que, si un chiffre est déclaré dans une section, un chiffre équivalent sera déclaré dans une autre section.

Vérification de la validité

Vérification qui porte sur un champ ou une cellule à la fois. Elle fait en sorte que nous tenions compte des descripteurs d'enregistrement, des caractères et des valeurs invalides; que les champs obligatoires aient été remplis (p.ex., aucun champ de quantité ne demeure vierge si un nombre est requis); que nous ayons bien utilisé les unités de mesure précisées et que le temps de déclaration respecte les limites précisées.

Vérification de l'étendue

Activité qui ressemble à la vérification de la validité, car elle porte aussi sur un champ à la fois. Ce type de vérification a pour but d'assurer que les valeurs, les rapports et les calculs respectent les limites prédéterminées.

Vérification des dédoublements

Activité par laquelle on examine un enregistrement complet à la fois. Elle recherche les dédoublements (c.-à-d. veille à ce que l'enregistrement d'une personne ou d'une chose apparaît une seule fois). Cette vérification recherche aussi si le répondant apparaît plus d'une fois dans l'univers de l'enquête, surtout après la modification du nom. Enfin, elle vérifie aussi si les données ont été entrées une seule fois dans le système.

Vérification des données

Activité par laquelle on s'assure que les données d'enquête sont exactes, exhaustives et cohérentes. (Voir aussi Imputation)

Vérifications diverses

Par vérifications diverses, on entend les dispositions spéciales de déclaration, les vérifications dynamiques propres à l'enquête, les vérifications de la classification appropriée, les modifications apportées aux adresses physiques, aux lieux et aux contacts ainsi que la vérification de la lisibilité, qui permet de s'assurer que les chiffres et les symboles sont reconnaissables.

Vérification historique

Activité par laquelle on compare les réponses des enquêtes régulières à celles des enquêtes précédentes. Par exemple, toute modification importante par rapport à la dernière enquête sera signalée. On compare aussi les rapports et les calculs. Ainsi, tout écart qui ne respecte pas les limites établies est noté et fait l'objet d'une analyse.

Vérification statistique

Vérification qui permet de comparer un ensemble complet de données. Ce type de vérification a lieu seulement après l'exécution de toutes les autres vérifications et la correction subséquente des données. Les données sont compilées et toute valeur extrême, toute donnée suspecte et toute valeur aberrante est rejetée.

X

$\overline{x}$

Symbole mathématique de la Moyenne.