1

DONNÉES ET STATISTIQUES

STATISTIQUES APPLIQUÉES

*Bloomberg Business Week**
New York, État de New York

Avec un tirage mondial de plus d'un million d'exemplaires, *Bloomberg Business Week* est le magazine d'information économique et financière le plus lu au monde. Les 1 700 reporters de Bloomberg, répartis dans 145 bureaux à travers le monde, sont en mesure de fournir une grande variété d'articles, suscitant l'intérêt des économistes et hommes d'affaires. En plus d'articles de fond traitant de sujets d'actualité, le magazine contient des articles relatifs au commerce international, à l'analyse économique, au traitement de l'information, aux sciences et technologies. Les informations contenues dans les articles de fond et les rubriques récurrentes aident les lecteurs à se tenir informés des développements récents dans les domaines considérés et à évaluer l'impact de ces derniers sur les affaires et les conditions économiques.

La plupart des numéros de *Bloomberg Business Week*, publiés auparavant sous le titre *Business Week*, contiennent un dossier détaillé sur un sujet d'actualité. Souvent, les dossiers détaillés contiennent des éléments et des résumés statistiques qui aident le lecteur à comprendre l'information économique. Par exemple, l'impact du développement du cloud computing sur les entreprises, la crise à laquelle fait face l'opérateur postal USPS ou les raisons qui font que la crise de la dette a été pire que prévue, ont fait l'objet de nombreux articles et de dossiers. De plus, *Bloomberg Business Week* fournit de nombreuses statistiques sur l'état de l'économie, dont des indices de production, le prix des actions, la valeur des fonds communs de placement et les taux d'intérêt.

Bloomberg Business Week utilise également des données et des informations statistiques pour gérer sa propre activité commerciale. Par exemple, une enquête annuelle auprès de ses abonnés aide la société à connaître leur profil, leurs habitudes de lecture, leurs achats, leur style de vie, etc. Les responsables de *Bloomberg Business Week* utilisent les résultats statistiques de l'enquête pour améliorer les services qu'ils offrent à leurs abonnés et aux annonceurs publicitaires. Une enquête récente a révélé que 90 % des abonnés Nord-Américains à *Bloomberg Business Week* utilisent un ordinateur personnel à la maison et que 64 % envisagent l'achat d'un ordinateur sur un plan professionnel. De telles statistiques avertissent les dirigeants de *Bloomberg Business Week* de l'intérêt que peuvent porter leurs abonnés à des articles relatifs aux nouveaux développements informatiques. De plus, les conclusions de ces enquêtes sont mises à la disposition d'annonceurs potentiels. Le pourcentage élevé d'abonnés utilisant un ordinateur à la maison et envisageant l'achat d'un ordinateur dans un cadre professionnel peut inciter certains fabricants à faire de la publicité pour leurs produits dans le magazine.

Dans ce chapitre, nous discuterons des types de données disponibles pour l'analyse statistique et décrirons les moyens de les obtenir. Nous introduirons ensuite les statistiques descriptives et l'inférence statistique en tant que moyens de convertir des données en information statistique utile et facilement interprétable.

---

* Les auteurs remercient Charlene Trentham, directrice de recherche, de leur avoir fourni ce Statistiques appliquées.

Fréquemment, on lit ce genre de phrases dans les journaux et les magazines :

- Le prix médian d'une maison individuelle ancienne s'élève à 186 000 dollars, en hausse de 7,6 % par rapport à l'an dernier (*The Wall Street Journal*, 8 novembre 2012).
- 14,1 % des directeurs généraux des sociétés appartenant au classement Fortune 500 sont des femmes (*The Wall Street Journal*, 30 avril 2012).
- Le coût annuel moyen d'une année d'étude s'élève à 17 100 dollars dans les universités publiques d'État et à 38 600 dollars dans les universités privées (*Money Magazine*, mars 2012).
- Une enquête de Yahoo Finance a révélé que 51 % des travailleurs pensent que la clé pour progresser réside dans la politique de promotion interne alors que 27 % pensent que la clé, c'est de travailler dur (*USA Today*, 29 septembre 2012).
- L'âge médian lors du premier mariage est de 29 ans pour les hommes et 26 ans pour les femmes (Associated Press, 25 décembre 2011).
- Le pourcentage de travailleurs américains dormant moins de six heures par nuit est de 30 % (*The Wall Street Journal*, 4 août 2012).
- Le découvert moyen des cartes de crédit est de 5 204 dollars par personne (site Internet de PRWeb, 5 avril 2012).

Les chiffres présents dans les phrases ci-dessus (186 000 dollars ; 7,6 % ; 14,1 % ; 17 100 dollars ; 38 600 dollars ; 51 % ; 27 % ; 29 ; 26 ; 30 % et 5 204 dollars) sont appelés statistiques. Ainsi, dans le langage courant, le terme « *statistique* » recouvre des données chiffrées telles que les moyennes, les médianes, les pourcentages et les valeurs maximales qui nous aident à comprendre l'environnement économique. Cependant, comme vous le verrez, le champ ou le contenu des statistiques inclut beaucoup plus que des chiffres. De façon plus générale, la **statistique** est l'art et la science de collecter, analyser, présenter et interpréter des données. Plus particulièrement en économie et dans le monde des affaires, l'information fournie par la collecte, l'analyse, la présentation et l'interprétation des données, offre aux dirigeants une meilleure compréhension de l'environnement économique et commercial et leur permet ainsi de prendre de bonnes décisions en toute connaissance de cause. Dans cet ouvrage, nous insistons sur l'utilisation des statistiques dans la prise de décision en matière économique et commerciale.

Le chapitre 1 débute par quelques exemples d'applications statistiques dans le monde des affaires et en économie. Dans la section 1.2, nous définissons le terme « *données* » et introduisons le concept d'ensemble de données. Cette section introduit également des termes clés comme « *variables* » et « *observations* », discute des différences entre données quantitatives et qualitatives et illustre l'utilisation des données en coupe transversale et les séries temporelles. La section 1.3 traite de la collecte des données à partir de sources existantes ou à partir d'enquêtes ou d'études expérimentales conçues pour obtenir de nouvelles données. Le rôle clé que joue désormais Internet dans la collecte de données est également souligné. L'utilisation des données pour développer des statistiques descriptives et faire de l'inférence statistique est décrite dans les sections 1.4 et 1.5. Les trois dernières sections du chapitre 1 décrivent le rôle de l'informatique dans l'analyse

statistique, fournissent une introduction au traitement des données et une discussion des bonnes pratiques statistiques. Une annexe à la fin du chapitre propose une introduction à l'outil statistique StatTools qui peut être utilisé pour élargir les possibilités d'analyse statistique offertes par Microsoft Excel.

1.1 APPLICATIONS EN ÉCONOMIE ET GESTION

Dans l'environnement économique et commercial actuel, tout le monde a accès à de nombreuses informations statistiques. Les dirigeants et les managers qui ont le plus de succès, sont ceux qui comprennent l'information et savent l'utiliser à bon escient. Dans cette section, nous présentons des exemples qui illustrent quelques utilisations de statistiques dans le domaine économique et commercial.

1.1.1 Comptabilité

Les experts comptables utilisent des procédures d'échantillonnage statistique lorsqu'ils effectuent des audits pour le compte de leurs clients. Par exemple, supposons qu'une entreprise de comptabilité veuille déterminer si le montant du compte « fournisseurs » qui apparaît dans le bilan, correspond bien au montant réel. Généralement, le nombre de fournisseurs est tellement grand que réexaminer et valider chaque compte individuellement serait trop long et trop coûteux. Dans de telles situations, il est courant que l'expert-comptable sélectionne un sous-ensemble de comptes, appelé échantillon. Après avoir réexaminé les comptes de l'échantillon, l'expert-comptable conclut si le montant du compte « fournisseurs » inscrit dans le bilan est acceptable ou non.

1.1.2 Finance

Les analystes financiers utilisent des informations statistiques diverses pour orienter leurs recommandations en matière d'investissement. Dans le cas de titres boursiers, les analystes examinent un certain nombre de données financières, telles que le coefficient de capitalisation des résultats et le rendement des dividendes. En comparant l'information pour un titre seul et l'information pour la moyenne des titres du marché, un analyste financier peut déjà savoir si le titre est un bon investissement. Par exemple, *The Wall Street Journal* (19 mars 2012) rapportait que le coefficient moyen de capitalisation des 500 sociétés formant l'indice S&P 500 était de 2,2 %. Le coefficient de capitalisation de Microsoft s'élevait à 2,42 %. Ces différentes informations statistiques sur le coefficient de capitalisation nous indiquent que le rendement de Microsoft était supérieur au rendement moyen des 500 sociétés composant l'indice S&P 500. Cette information, ajoutée à d'autres, pourrait aider l'analyste financier à recommander l'achat, la vente ou la conservation des actions Microsoft.

1.1.3 Marketing

Les scanners électroniques des caisses enregistreuses dans les commerces collectent des données, utilisées dans de nombreuses applications de recherche en marketing. Par exemple, des sociétés telles que ACNielsen et Information Resources achètent les données recueillies par les scanners des caisses enregistreuses, les exploitent et vendent ensuite les conclusions statistiques aux fabricants. Les fabricants dépensent des centaines de milliers de dollars par catégorie de produit pour obtenir ce type de données scannées. Ils achètent également les données et les conclusions statistiques relatives aux activités promotionnelles, telles que les offres spéciales en tête de gondole dans les magasins. Les responsables de la marque peuvent examiner les conclusions des études statistiques menées à partir des données scannées afin de mieux comprendre la relation entre vente et promotion. De telles analyses se révèlent souvent utiles pour établir les futures stratégies commerciales des produits concernés.

1.1.4 Production

L'importance accordée de nos jours à la qualité fait de son contrôle une application primordiale de la statistique, dans la gestion de la production. De nombreux graphiques de contrôle de la qualité sont utilisés pour vérifier les caractéristiques du produit fini dans un processus de production. En particulier, un diagramme en barres peut être utilisé pour contrôler la production moyenne. Supposons, par exemple, qu'une machine remplisse des canettes de 33 cl d'une boisson non-alcoolisée. Périodiquement, un agent de production sélectionne un échantillon de canettes et calcule la quantité moyenne contenue dans les canettes de l'échantillon. Cette moyenne, ou valeur $\overline{x}$, est représentée sur un graphique de la moyenne. Un point situé au-dessus de la limite supérieure du graphique indique un sur-remplissage alors qu'un point situé en-dessous de la limite inférieure indique un sous-remplissage. Le processus de production est dit « sous contrôle » et peut se poursuivre tant que les points représentés sur le graphique de la moyenne sont compris entre les limites inférieure et supérieure. L'interprétation correcte d'un diagramme en barres permet de déterminer si des ajustements sont nécessaires, afin de corriger le processus de production.

1.1.5 Économie

Les économistes fournissent fréquemment des prévisions à propos de certains faits économiques futurs. Ils utilisent de nombreuses informations statistiques pour effectuer ces prévisions. Par exemple, pour prévoir le taux d'inflation, les économistes utilisent des indicateurs tels que l'indice des prix à la production, le taux de chômage et le taux d'utilisation des capacités de production. Souvent, ces indicateurs statistiques sont intégrés à des modèles de prévision qui prévoient le taux d'inflation.

1.1.6 Les systèmes d'information

Les administrateurs des systèmes d'information sont responsables au jour le jour du fonctionnement des réseaux informatiques de l'entreprise. Une grande quantité d'information

statistique permet aux administrateurs réseaux d'évaluer la performance des outils informatiques, des réseaux locaux ou à distance, de l'intranet et des autres moyens de communication. Des statistiques telles que le nombre moyen d'utilisateurs du système, la durée durant laquelle chaque composant du système n'est pas utilisé et la part de la bande passante utilisée à différents moments de la journée sont des exemples d'informations statistiques qui aident l'administrateur des systèmes informatiques à mieux comprendre et gérer le réseau informatique.

Les applications statistiques telles que celles décrites dans cette section font partie intégrante de cet ouvrage. De tels exemples fournissent un aperçu de l'étendue des applications statistiques. Pour compléter ces exemples, nous avons demandé à des personnes utilisant des statistiques dans les domaines commercial et économique, de rédiger des articles dans la section intitulée « Statistiques appliquées », afin d'introduire les outils présentés dans chaque chapitre. Les applications décrites dans Statistiques appliquées illustrent concrètement l'importance des statistiques.

1.2 DONNÉES

Les *données* sont les faits et les chiffres qui sont collectés, analysés et résumés pour pouvoir ensuite être interprétés. Toutes les données collectées dans une étude particulière forment **l'ensemble de données** de l'étude. Le tableau 1.1 présente un ensemble de données contenant des informations relatives à 60 pays qui font partie de l'Organisation mondiale du commerce. L'Organisation mondiale du commerce encourage le libre-échange au niveau international et constitue une plateforme de résolution des conflits commerciaux.

Tableau 1.1 ***Ensemble de données pour les 60 pays de l'Organisation mondiale du commerce***

Pays	Statut à l'OMC	PIB par tête ($)	Déficit de la balance commerciale (en milliers de $)	Note Fitch	Perspective Fitch
Arménie	Membre	5 400	2 673 359	BB–	Stable
Australie	Membre	40 800	–33 304 157	AAA	Stable
Autriche	Membre	41 700	12 796 558	AAA	Stable
Azerbaïdjan	Observateur	5 400	–16 747 320	BBB–	Positive
Bahreïn	Membre	27 300	3 102 665	BBB	Stable
Belgique	Membre	37 600	–14 930 833	AA+	Negative
Brésil	Membre	11 600	–29 796 166	BBB	Stable
Bulgarie	Membre	13 500	4 049 237	BBB–	Positive
Canada	Membre	40 300	–1 611 380	AAA	Stable
Cap Vert	Membre	4 000	874 459	B+	Stable
Chili	Membre	16 100	–14 558 218	A1	Stable

Chine	Membre	8 400	–156 705 311	A1	Stable
Colombie	Membre	10 100	–1 561 199	BBB–	Stable
Costa Rica	Membre	11 500	5 807 509	BB+	Stable
Croatie	Membre	18 300	8 108 103	BBB–	Negative
Chypre	Membre	29 100	6 623 337	BBB	Negative
République tchèque	Membre	25 900	–10 749 467	A+	Positive
Danemark	Membre	40 200	–15 057 343	AAA	Stable
République de l'Équateur	Membre	8 300	1 993 819	B–	Stable
Égypte	Membre	6 500	28 486 933	BB	Negative
Salvador	Membre	7 600	5 019 363	BB	Stable
Estonie	Membre	20 200	802 234	A+	Stable
France	Membre	35 000	118 841 542	AAA	Stable
Géorgie	Membre	5 400	4 398 153	B+	Positive
Allemagne	Membre	37 900	–213 367 685	AAA	Stable
Hongrie	Membre	19 600	–9 421 301	BBB–	Negative
Islande	Membre	38 000	–504 939	BB+	Stable
Irlande	Membre	39 500	–59 093 323	BBB+	Negative
Israël	Membre	31 000	6 722 291	A	Stable
Italie	Membre	30 100	33 568 668	A+	Negative
Japon	Membre	34 300	31 675 424	AA	Negative
Kazakhstan	Observateur	13 000	–33 220 437	BBB	Positive
Kenya	Membre	1 700	9 174 198	B+	Stable
Lettonie	Membre	15 400	2 448 053	BBB–	Positive
Liban	Observateur	15 600	13 715 550	B	Stable
Lituanie	Membre	18 700	3 359 641	BBB	Positive
Malaisie	Membre	15 600	–39 420 064	A–	Stable
Mexique	Membre	15 100	1 288 112	BBB	Stable
Pérou	Membre	10 000	–7 888 993	BBB	Stable
Philippines	Membre	4 100	15 667 209	BB+	Stable
Pologne	Membre	20 100	19 552 976	A–	Stable
Portugal	Membre	23 200	21 060 508	BBB–	Negative
Corée du Sud	Membre	31 700	–37 509 141	A+	Stable
Roumanie	Membre	12 300	13 323 709	BBB–	Stable
Russie	Observateur	16 700	–151 400 000	BBB	Positive
Rwanda	Membre	1 300	939 222	B	Stable
Serbie	Observateur	10 700	8 275 693	BB–	Stable
Seychelles	Observateur	24 700	666 026	B	Stable
Singapour	Membre	59 900	–27 110 421	AAA	Stable
Slovaquie	Membre	23 400	–2 110 626	A+	Stable
Slovénie	Membre	29 100	2 310 617	AA–	Negative
Afrique du Sud	Membre	11 000	3 321 801	BBB+	Stable
Suède	Membre	40 600	–10 903 251	AAA	Stable
Suisse	Membre	43 400	–27 197 873	AAA	Stable

Thaïlande	Membre	9 700	2 049 669	BBB	Stable
Turquie	Membre	14 600	71 612 947	BB+	Positive
Royaume-Uni	Membre	35 900	162 316 831	AAA	Negative
Uruguay	Membre	15 400	2 662 628	BB	Positive
États-Unis	Membre	48 100	784 438 559	AAA	Stable
Zambie	Membre	1 600	–1 805 198	B+	Stable

1.2.1 Éléments, variables et observations

Les **éléments** sont les entités auprès desquelles les données sont collectées. Chaque pays listé dans le tableau 1.1 est un élément, dont le nom apparaît dans la première colonne. Puisqu'il y a 60 pays, l'ensemble de données contient 60 éléments.

Une **variable** est une caractéristique des éléments à laquelle on s'intéresse. L'ensemble de données du tableau 1.1 contient les cinq variables suivantes :

- Le statut à l'OMC : le statut de membre du pays au sein de l'Organisation mondiale du commerce ; le pays peut être membre ou observateur.
- Le PIB par tête ($) : la production globale du pays divisée par le nombre d'habitants du pays ; il s'agit d'une variable communément utilisée pour comparer la productivité économique des pays.
- Le déficit de la balance commerciale (en milliers de dollars) : la différence entre la valeur (en dollars) des importations et des exportations du pays.
- La note Fitch : l'évaluation de la dette souveraine du pays établie par le groupe Fitch[1] ; les notes vont de AAA à F et peuvent être modulées par + ou -.
- Les perspectives Fitch : un indicateur de la tendance vers laquelle la note pourrait tendre dans les deux ans à venir ; les prévisions peuvent être négatives, stables ou positives.

Les données sont obtenues en collectant des informations sur chaque variable pour tous les éléments de l'étude. L'ensemble des informations obtenues pour un élément particulier correspond à une **observation**. En se référant au tableau 1.1, nous voyons que la première observation contient l'ensemble des informations suivantes : Membre, 5 400, 2 673 359, BB- et Stable. La seconde contient les informations suivantes : Membre, 40 800, -33 304 157, AAA et Stable ; et ainsi de suite. Un ensemble de données de 60 éléments contient 60 observations.

1.2.2 Échelles de mesure

Différentes échelles de mesure d'une variable existent : nominale, ordinale, par intervalle (ou cardinale) ou de rapport. L'échelle de mesure détermine la quantité d'information contenue dans les données et indique la méthode d'analyse des données la plus appropriée.

1 Le groupe Fitch est l'une des trois institutions de notation reconnues aux États-Unis, certifiées par la Commission de contrôle des marchés financiers américaine, la SEC (Securities and Exchanges Commission). Les deux autres sont Standard and Poor's et Moody's.

Lorsque les données d'une variable consistent en des labels ou des noms utilisés pour identifier une caractéristique de l'élément, l'échelle de mesure est **nominale**. Par exemple, en se référant au tableau 1.1, nous voyons que l'échelle de mesure de la variable « Statut à l'OMC » est nominale, les qualitatifs « membre » ou « observateur » étant utilisés pour identifier le statut du pays au sein de l'OMC. Dans les cas où l'échelle de mesure est nominale, un code numérique ou alpha-numérique peut être utilisé. Par exemple, pour faciliter la collecte de données et préparer les données en vue de leur incorporation dans une base de données informatisée, nous pourrions utiliser un code numérique, en attribuant le chiffre 1 aux pays membres, le chiffre 2 aux pays observateurs. L'échelle de mesure est nominale même si les données apparaissent sous la forme de valeurs numériques.

L'échelle de mesure d'une variable est **ordinale** si les données exhibent les propriétés nominales et qu'il est possible de les ordonner (si cela a un sens). Par exemple, en se référant aux données du tableau 1.1, l'échelle de mesure pour la note Fitch est ordinale puisque les notes qui vont de AAA à F, peuvent être ordonnées de la meilleure à la moins bonne note. Le système de notation par lettre possède les propriétés des données nominales mais en plus, ces données peuvent être classées ou ordonnées, ce qui implique que l'échelle de mesure est ordinale. Les données ordinales peuvent également être enregistrées sous forme de code numérique, par exemple, votre classement à l'école.

L'échelle de mesure d'une variable devient **cardinale** (ou **par intervalle**) si les données possèdent les propriétés ordinales et si l'intervalle entre les valeurs peut être mesuré par une unité de mesure fixe. Les données cardinales (ou par intervalle) sont toujours numériques. Les résultats d'un test d'aptitude intellectuelle sont un exemple de données cardinales. Par exemple, les résultats de trois étudiants à un test de mathématiques (620, 550 et 470) peuvent être ordonnés de la meilleure à la moins bonne performance. De plus, les écarts entre les résultats ont un sens. Par exemple, l'étudiant 1 a obtenu 620-550 = 70 points de plus que l'étudiant 2, alors que l'étudiant 2 a obtenu 550-470 = 80 points de plus que l'étudiant 3.

L'échelle de mesure d'une variable est dite **de rapport** si les données ont toutes les propriétés des données cardinales et que le rapport entre deux valeurs a un sens. Des variables telles que la distance, la hauteur, le poids et la durée, utilisent une échelle de rapport. Cette échelle nécessite l'inclusion d'une valeur nulle pour indiquer que rien n'existe pour la variable au point zéro. Par exemple, considérons le coût d'une automobile. Une valeur nulle indique que l'automobile a un coût nul et est gratuite. De plus, si nous comparons une automobile dont le coût est de 30 000 dollars à une autre automobile dont le coût est de 15 000 dollars, le rapport indique que le coût de la première automobile est deux fois plus élevé que celui de la seconde.

1.2.3 Données qualitatives et données quantitatives

Par ailleurs, les données peuvent être classées en fonction de leur nature qualitative ou quantitative. Les données qui peuvent être regroupées par catégorie sont des **données qualitatives (ou catégorielles).** L'échelle de mesure des données qualitatives peut être ordinale ou nominale. Les données qui prennent des valeurs numériques pour indiquer des

quantités sont des **données** dites **quantitatives**. Les données quantitatives ont une échelle de mesure cardinale ou de rapport.

Une **variable qualitative (ou catégorielle)** est une variable dont les données sont qualitatives, et une **variable quantitative** est une variable dont les données sont quantitatives. L'analyse statistique appropriée à une variable particulière dépend de sa nature qualitative ou quantitative. Si la variable est qualitative, l'analyse statistique est plutôt limitée. Nous pouvons résumer des données qualitatives en dénombrant le nombre d'observations ou en calculant la proportion d'observations dans chaque catégorie. Cependant, même lorsque des données qualitatives sont identifiées par un code numérique, des opérations arithmétiques telles que l'addition, la soustraction, la multiplication et la division, ne permettent pas d'obtenir des résultats ayant un sens. La section 2.1 traite des méthodes d'analyse des données qualitatives.

La méthode statistique appropriée pour résumer des données dépend de la nature quantitative ou qualitative des données.

Par contre, les opérations arithmétiques fournissent des résultats ayant un sens lorsque les variables sont quantitatives. Par exemple, des données quantitatives peuvent être additionnées et divisées par le nombre d'observations de façon à calculer la valeur moyenne. Cette moyenne a un sens mathématique et est facilement interprétable. En général, les outils d'analyse statistique sont plus nombreux pour des données quantitatives. La section 2.2 et le chapitre 3 présentent les méthodes d'analyse statistique des données quantitatives.

1.2.4 Données en coupe transversale et séries temporelles

Pour les besoins de l'analyse statistique, la distinction entre les données en coupe transversale et les séries temporelles est fondamentale. Les **données en coupe transversale** sont collectées au même moment (ou approximativement au même moment). Les données du tableau 1.1 sont en coupe transversale puisqu'elles décrivent les cinq variables pour les 60 nations de l'Organisation mondiale du commerce à un même moment dans le temps. Les **séries temporelles** sont des données collectées sur plusieurs périodes de temps différentes. Par exemple, la figure 1.1 représente le prix moyen d'un gallon d'essence sans plomb aux États-Unis entre 2007 et 2012. Notez que le prix de l'essence sans plomb a atteint un point haut durant l'été 2008 puis a fortement chuté durant l'automne 2008. Depuis 2008, le prix moyen d'un gallon d'essence a régulièrement augmenté, approchant d'un nouveau sommet en 2012.

On trouve fréquemment dans les publications économiques une représentation graphique des séries temporelles. De tels graphiques aident les analystes à comprendre ce qui s'est passé, à identifier les tendances au cours du temps et à prévoir les niveaux futurs des séries temporelles. On trouve diverses formes de graphiques de séries temporelles, comme illustré par la figure 1.2. Avec quelques connaissances, ces graphiques sont généralement faciles à comprendre et interpréter. Par exemple, le graphique A sur la figure 1.2

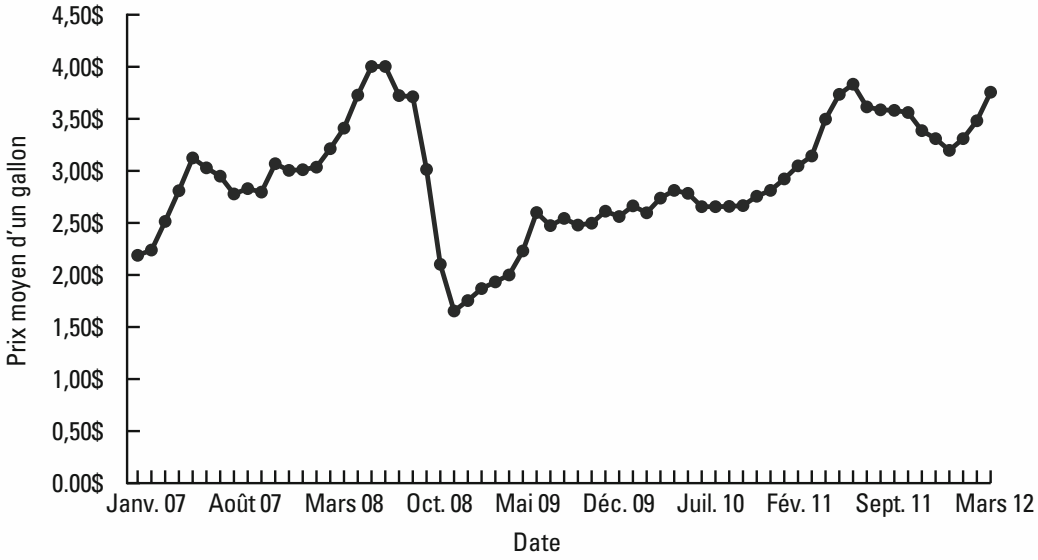


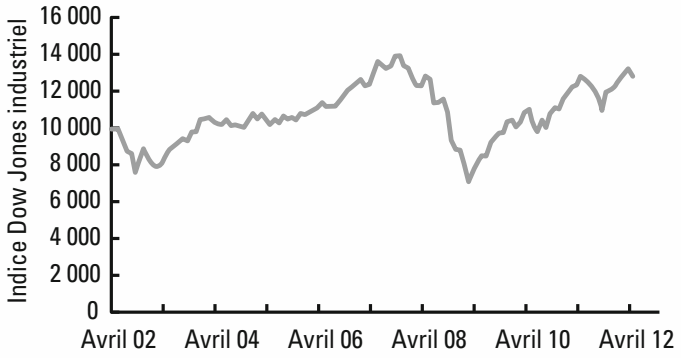
Figure 1.1 ***Prix moyen d'un gallon d'essence sans plomb aux États-Unis***

Source : Administration américaine de l'information sur l'énergie, Département américain de l'énergie, mars 2012.

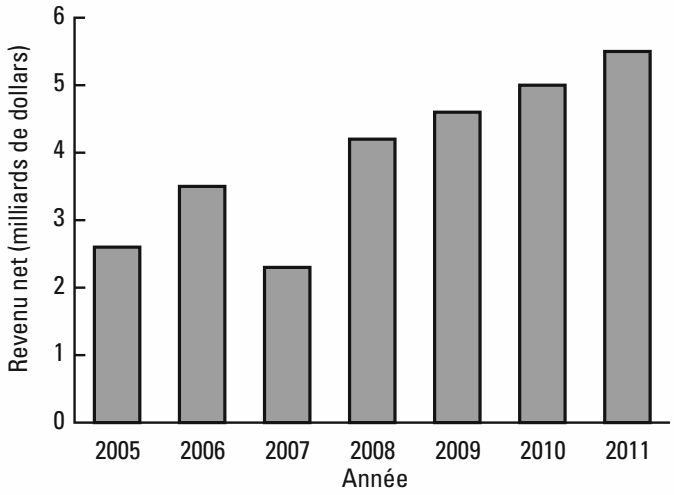
représente l'indice Dow Jones Industriel de 2002 à 2012. En avril 2002, l'indice était proche de 10 000 points. Au cours des cinq années suivantes, l'indice a augmenté jusqu'à son plus haut niveau jamais atteint, plus de 14 000 points en octobre 2007. Cependant, notez la chute brutale de l'indice après ce record de 2007. En mars 2009, l'indice était revenu à 7 000 points en raison d'un contexte économique défavorable. Ce fut une période effrayante et décourageante pour les investisseurs. Toutefois, fin 2009, l'indice a commencé à se redresser, atteignant 10 000 points. Il a régulièrement progressé ensuite et était supérieur à 13 000 points début 2012.

Le graphique B représente le revenu net de la société McDonald's entre 2005 et 2011. La crise économique de 2008 et 2009 fut plutôt bénéfique à MacDonald's, son revenu net atteignant un record historique. La croissance du revenu net de la société illustre le fait que la société a prospéré durant la crise : les ménages ont réduit leurs dépenses en fréquentant moins les restaurants plus chers et en se rabattant sur les alternatives moins onéreuses offertes par McDonald's. Le revenu net de McDonald's a continué à progresser, atteignant des niveaux jamais atteints en 2010 et 2011.

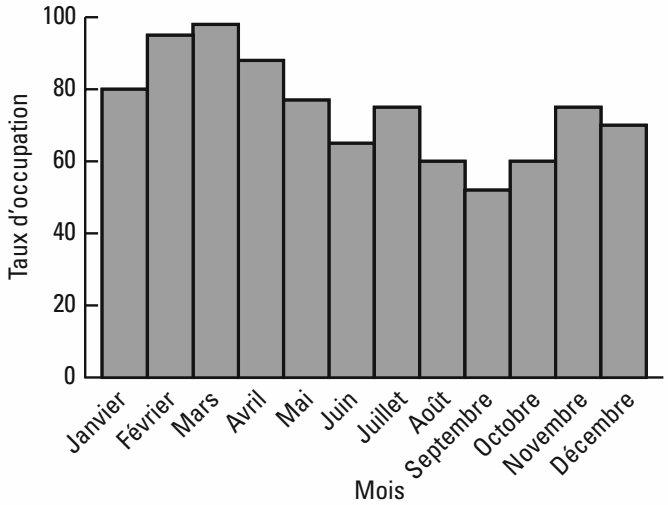
Le graphique C illustre une série temporelle des taux d'occupation des hôtels dans le Sud de la Floride au cours d'une année. Les taux d'occupation les plus élevés entre 95 % et 98 % sont observés durant les mois de février et mars lorsque le climat du Sud de la Floride est le plus attractif pour les touristes. En réalité, la saison haute pour les



(A) Indice Dow Jones industriel



(B) Revenu net de la société McDonalds



(C) Taux d'occupation des hôtels du Sud de la Floride

Figure 1.2 ***Quelques représentations graphiques de séries temporelles***

hôteliers du Sud de la Floride s'étend généralement du mois de janvier au mois d'avril. D'un autre côté, observez les taux d'occupation d'août à octobre : le taux d'occupation le plus faible (50 %) est observé en septembre. Les températures élevées et la saison des ouragans expliquent cette baisse de la fréquentation des hôtels en cette période.

REMARQUES

1. Une observation est un ensemble de mesures obtenues pour chaque élément d'un ensemble de données. Ainsi, le nombre d'observations et le nombre d'éléments sont identiques. Le nombre de mesures obtenues sur chaque élément est égal au nombre de variables. Par conséquent, le nombre total de valeurs dans un ensemble de données peut être obtenu en multipliant le nombre d'observations par le nombre de variables.
2. Les données quantitatives peuvent être discrètes ou continues. Celles qui mesurent une variable dénombrable (par exemple, le nombre d'appels reçus en 5 minutes) sont discrètes. Celles qui mesurent des variables indénombrables (par exemple, le poids ou le temps) sont continues, aucune séparation n'étant possible entre les valeurs potentielles des données.

1.3 SOURCES DE DONNÉES

Les données peuvent être obtenues à partir de sources existantes ou grâce à des enquêtes ou des études menées spécifiquement dans le but de collecter de nouvelles données.

1.3.1 Sources existantes

Dans certains cas, les données nécessaires à une application particulière existent déjà. De nombreuses entreprises constituent des bases de données sur leurs employés, leurs clients et leurs opérations commerciales. Des données sur le salaire, l'âge et les années de service des employés peuvent généralement être obtenues auprès du service du personnel. D'autres services internes à l'entreprise collectent des données sur les ventes, les dépenses publicitaires, les coûts de distribution, l'inventaire et les quantités produites. La plupart des entreprises entretiennent également des bases de données sur leurs clients. Le tableau 1.2 fournit quelques exemples de données fréquemment disponibles dans les services internes des entreprises.

Des organismes spécialisés dans la collecte et le traitement des données fournissent des quantités substantielles de données économiques et commerciales. Les entreprises ont accès à ces sources de données externes par des arrangements de crédit-bail ou par achat. Dun & Bradstreet, Bloomberg et Dow Jones & Company sont trois entreprises qui fournissent de vastes services en matière de collecte de données. Les sociétés

Tableau 1.2 *Exemples de données disponibles dans les registres internes de l'entreprise*

Source	Types de données disponibles
Registre des employés	Nom, adresse, numéro de sécurité sociale, salaire, nombre de jours de congé, nombre de jours d'arrêt maladie et prime.
Registre de la production	Référence de la pièce ou du produit, quantité produite, coût direct du travail et coût des matériaux.
Inventaire	Référence de la pièce ou du produit, nombre d'unités disponibles, prévision de production, quantité commandée et grille tarifaire.
Registre des ventes	Référence du produit, volume des ventes, volume des ventes par région et par type de client.
Registre des crédits	Nom du client, adresse, numéro de téléphone, crédit maximal et solde des créances.
Profil des clients	Âge, sexe, niveau de revenu, taille du ménage, adresse et préférences.

ACNielsen et Information Resources prospèrent grâce à la collecte et au traitement des données, qu'elles vendent ensuite à des annonceurs et à des producteurs.

De nombreuses associations industrielles et organisations de lobbying disposent également de nombreuses données. L'association américaine de l'industrie du tourisme conserve des informations relatives au tourisme, comme le nombre de touristes et le montant des dépenses touristiques par État. De telles informations peuvent intéresser l'industrie du tourisme. Le conseil d'admission des écoles supérieures de commerce conserve des données sur les résultats des tests, les caractéristiques des étudiants et le programme des cours. La plupart des données issues de ces sources sont accessibles à un coût modeste.

Internet est une source importante de données et d'informations statistiques. La plupart des sociétés possèdent leur site Web, sur lequel apparaissent des informations générales sur la société, ainsi que des données sur les ventes, le nombre d'employés, la gamme de produits, leurs prix et leurs spécificités. De plus, certaines entreprises se sont désormais spécialisées dans la divulgation d'informations sur Internet. En conséquence, tout le monde peut obtenir les cotations boursières, les prix d'un repas au restaurant, des données sur les salaires et une quantité d'informations quasi infinie.

Tableau 1.3 *Exemples de données disponibles auprès de quelques agences gouvernementales*

Agence gouvernementale	Données disponibles
Bureau des recensements	Données sur la population, le nombre de ménages et leurs revenus.
Banque centrale américaine	Données sur l'offre de monnaie, le crédit, le taux de change et le taux d'escompte.
Ministère des finances	Données sur le revenu, les dépenses et la dette du gouvernement fédéral.
Département du commerce	Données sur l'activité commerciale, la valeur des ventes par industrie, le niveau de profit par industrie, les industries en déclin et en croissance.
Bureau des statistiques du travail	Dépenses des ménages, salaires horaires, taux de chômage, sécurité au travail, statistiques internationales.

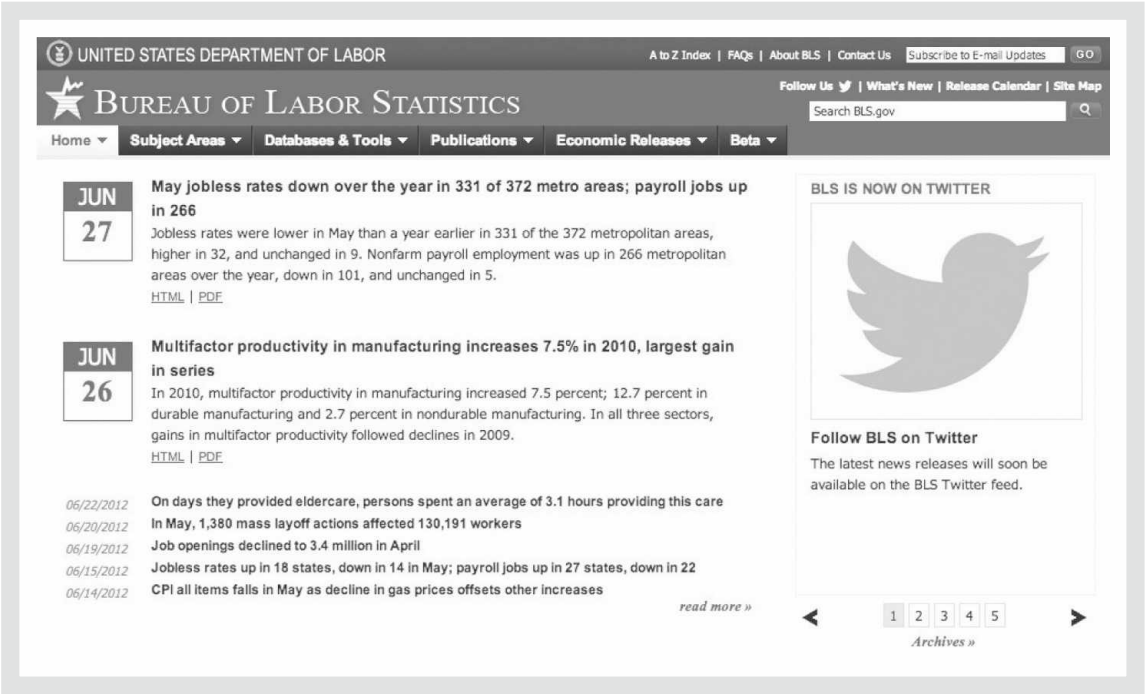


Figure 1.3 ***La page d'accueil du site Internet du bureau américain des statistiques du travail***

Les agences gouvernementales sont une autre source importante de données existantes. Par exemple, le département américain du travail conserve des données sur le taux d'embauche, les salaires, la taille de la population active et le degré de syndicalisation. Le tableau 1.3 fournit la liste de quelques agences gouvernementales et des données dont elles disposent. La plupart des agences gouvernementales qui collectent et traitent des données, rendent également public le résultat de leurs investigations sur un site Internet. La figure 1.3 présente la page d'accueil du site Internet du bureau américain des statistiques du travail.

1.4 ÉTUDES STATISTIQUES

Parfois les données nécessaires à une étude particulière ne sont pas disponibles auprès de sources existantes. Dans ces cas, les données peuvent être obtenues en effectuant une étude statistique. On distingue deux types d'études statistiques : les **études expérimentales** et les **études empiriques**.

La plus importante étude statistique expérimentale jamais réalisée est, semble-t-il, l'expérience réalisée par le Service public de la santé en 1954 relative à la campagne de vaccination contre la polio. Près de deux millions d'enfants scolarisés dans le primaire ont été sélectionnés à travers les États-Unis.

Dans une étude expérimentale, on identifie en premier lieu la variable qui nous intéresse. Ensuite, une ou plusieurs autres variables sont identifiées et contrôlées de sorte à obtenir des informations sur leur influence sur la variable d'intérêt. Prenons l'exemple d'une entreprise pharmaceutique intéressée par une étude permettant de connaître l'effet d'un nouveau médicament sur la pression artérielle. La pression artérielle est la variable d'intérêt de l'étude. Le dosage du nouveau médicament est une autre variable, supposée avoir un effet sur la pression artérielle. Pour obtenir des données concernant l'effet de ce nouveau médicament, les chercheurs sélectionnent un échantillon d'individus. Le dosage du nouveau médicament est contrôlé : chaque groupe d'individus reçoit un dosage différent. Les données sur la pression artérielle, avant et après traitement, sont collectées pour

Date : Nom du serveur :

Nos clients sont notre première priorité. Veuillez s'il vous plaît prendre quelques instants pour renseigner ce questionnaire, afin de nous permettre de mieux répondre à vos souhaits. Vous pouvez remettre cette carte à notre hôtesse en sortant ou la renvoyer par courrier électronique. Merci.

Service concerné	Excellent	Bon	Satisfaisant	Insatisfaisant
Qualité globale	❑	❑	❑	❑
Accueil par le maître d'hôtel	❑	❑	❑	❑
Déroulement du service	❑	❑	❑	❑
Service global	❑	❑	❑	❑
Professionnalisme	❑	❑	❑	❑
Connaissance du menu	❑	❑	❑	❑
Gentillesse	❑	❑	❑	❑
Sélection de vins	❑	❑	❑	❑
Sélection des menus	❑	❑	❑	❑
Qualité des plats	❑	❑	❑	❑
Présentation des plats	❑	❑	❑	❑
Rapport qualité-prix	❑	❑	❑	❑

Quels commentaires pouvez-vous faire pour nous aider à améliorer notre service ?

Merci, nous apprécions vos commentaires. L'équipe du Chops City Grill.

Figure 1.4 ***Sondage d'opinion auprès des clients du restaurant Chops City Grill de Naples, dans l'État de Floride***

chaque groupe. L'analyse statistique des données expérimentales permettra de déterminer l'influence du nouveau médicament sur la pression artérielle.

> Les études sur les fumeurs et les non-fumeurs sont des études empiriques puisque les chercheurs ne déterminent ou ne contrôlent pas qui fume et qui ne fume pas.

Les études statistiques non-expérimentales, ou empiriques, ne tentent pas de contrôler les variables d'intérêt. Un sondage est le type le plus courant d'études empiriques. Par exemple, lors d'un sondage en face-à-face, on identifie d'abord les questions. Ensuite un questionnaire est établi et distribué à un échantillon d'individus. Certains restaurants utilisent des études empiriques pour connaître l'opinion de leurs clients sur la qualité des menus, du service, de l'ambiance, etc. La figure 1.4 présente le questionnaire utilisé par le restaurant Chops City Grill de Naples, en Floride. Les clients interrogés doivent évaluer 12 variables : la qualité globale, l'accueil par le maître d'hôtel, le service, etc. Les catégories de réponse – excellent, bon, moyen, satisfaisant et insatisfaisant – permettent aux propriétaires du Chops City Grill de maintenir un haut niveau de qualité des plats proposés et du service.

Quiconque désire utiliser des données et des analyses statistiques en tant qu'outil d'aide à la décision, doit être conscient du coût et du temps que nécessite l'obtention des données. L'utilisation de sources existantes est souhaitable lorsque les données doivent être obtenues rapidement. Si les données importantes ne sont pas disponibles auprès d'une source existante, le temps et les coûts d'acquisition des données doivent être évalués. Dans tous les cas, il est important de considérer la contribution de l'analyse statistique dans le processus de prise de décision. Le coût d'acquisition des données et de l'analyse qui en découle, ne doit pas excéder les gains générés par l'utilisation de l'information pour prendre une meilleure décision.

1.4.1 Erreurs dans la collecte des données

Il convient de toujours avoir à l'esprit que des erreurs peuvent être commises lors de la collecte des données. Utiliser des données erronées peut s'avérer pire que de ne pas en utiliser. Une erreur dans l'acquisition des données intervient lorsque la valeur inscrite ne correspond pas à la vraie valeur, c'est-à-dire celle qui aurait été obtenue avec une procédure d'acquisition correcte. De telles erreurs peuvent survenir de différentes manières. Par exemple, un enquêteur peut faire une erreur d'enregistrement, et enregistrer 42 ans au lieu de 24 ans, ou bien la personne interrogée peut mal interpréter la question et donner une réponse incorrecte.

Les analystes expérimentés prennent grand soin de ne pas faire d'erreurs dans la collecte et l'enregistrement des données. Des procédures de détection des incohérences existent. Par exemple, l'attention de l'analyste est attirée lorsque le traitement d'un questionnaire révèle qu'un individu âgé de 22 ans a 20 années d'expérience professionnelle. Les analystes réexaminent également les données pour lesquelles on constate des valeurs inhabituellement élevées ou faibles, pouvant être dues à des erreurs de collecte. Dans le chapitre 3, nous présenterons quelques méthodes utilisées par les statisticiens, pour identifier ces valeurs « aberrantes ».

Les erreurs surviennent souvent au cours de la phase de collecte des données. Utiliser toutes les données disponibles de façon aveugle ou utiliser des données qui n'ont pas fait l'objet de toutes les attentions peut apporter une information trompeuse et conduire à prendre de mauvaises décisions. Ainsi, en prenant soin de collecter des données précises, on améliore le processus décisionnel.

1.5 STATISTIQUES DESCRIPTIVES

La plupart des informations statistiques contenues dans les journaux, les magazines, les rapports d'activité de sociétés et autres publications sont des données résumées et présentées sous une forme facilement compréhensible par le lecteur. On appelle de tels résumés sous forme de tableaux, de graphiques ou sous forme numérique, des **statistiques descriptives**.

On se réfère une fois encore à l'ensemble de données relatif aux 60 pays de l'Organisation mondiale du commerce, présenté dans le tableau 1.1. Des statistiques descriptives peuvent être utilisées pour résumer ces données. Par exemple, considérez la variable « Perspective Fitch » qui indique la direction dans laquelle la note du pays pourrait évoluer au cours des deux prochaines années. La perspective Fitch peut être négative, stable ou positive. Le tableau 1.4 présente un résumé sous forme de tableau des données indiquant, pour chaque tendance possible, le nombre pays présentant cette perspective. La figure 1.5 est un résumé graphique de ces mêmes données, sous forme d'un diagramme en barres. Ces types de représentations graphiques et sous forme de tableaux facilitent l'interprétation des données. En se référant au tableau 1.4 et à la figure 1.5, on s'aperçoit que la majorité des notes devraient être stables, 65 % des pays ayant une perspective d'évolution stable de leur note établie par Fitch. Les proportions de perspectives négatives et positives sont similaires, avec légèrement plus de pays ayant une perspective négative (18,3 %) qu'une perspective positive (16,7 %).

La figure 1.6 est un résumé graphique des données de la variable quantitative PIB par tête figurant dans le tableau 1.1, sous la forme d'un histogramme. À partir de cet histogramme, il est facile de voir que le PIB par tête des 60 pays est compris entre 0 et 60 000 dollars, les plus fortes concentrations se situant entre 10 000 et 20 000 dollars. Un seul pays a un PIB par tête supérieur à 50 000 dollars.

Tableau 1.4 *Fréquences et fréquences en pourcentage de la perspective d'évolution de la note Fitch des 60 pays*

Perspective Fitch	Fréquence	Fréquence en pourcentage
Positive	10	16,7
Stable	39	65,0
Négative	11	18,3

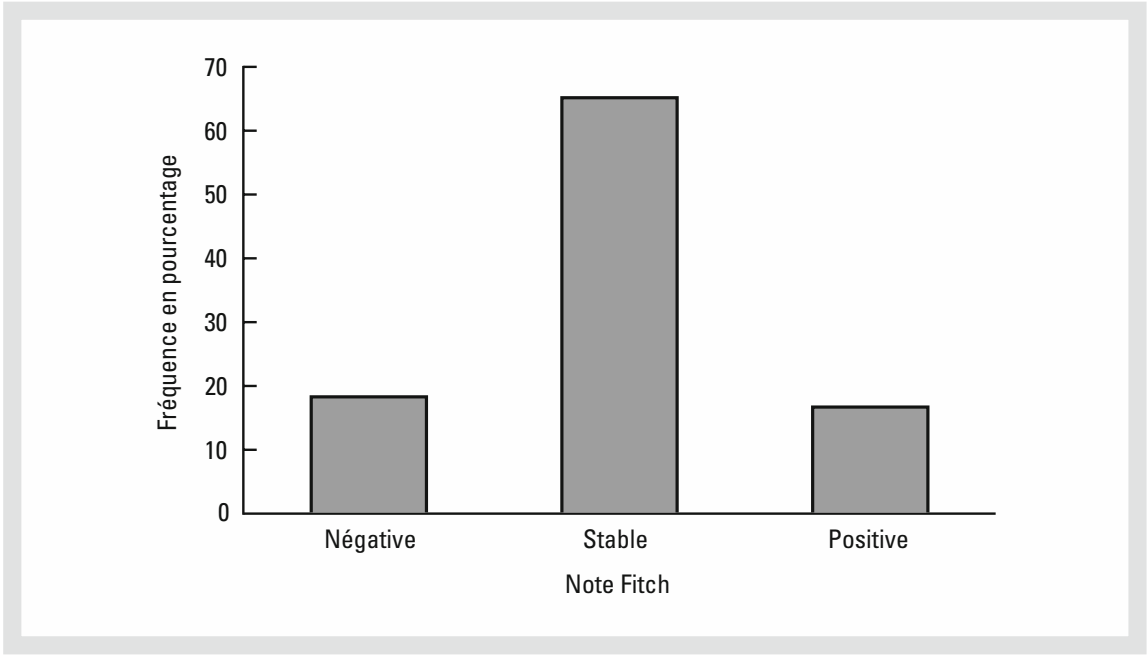


Figure 1.5 ***Diagramme en barres de la perspective d'évolution de la note Fitch des 60 pays***

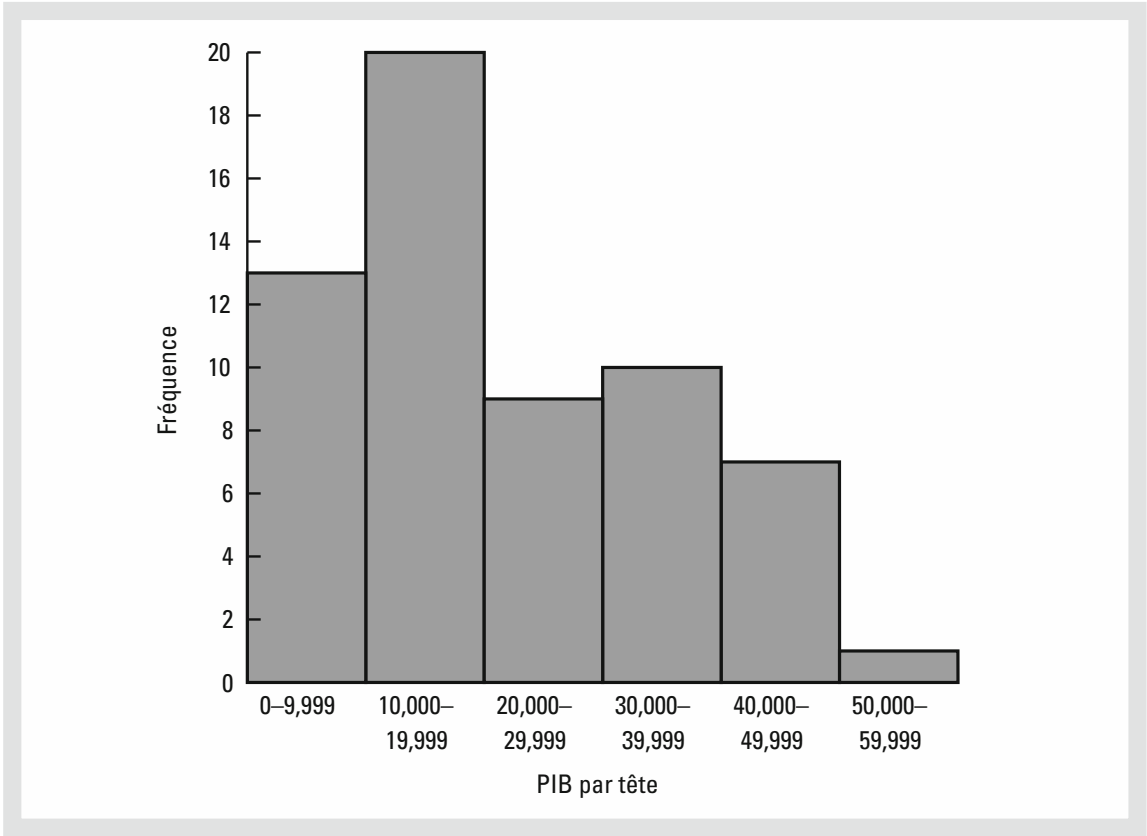


Figure 1.6 ***Histogramme du PIB par tête des 60 pays***

En plus des présentations sous forme de tableaux et de graphiques, on peut utiliser des statistiques descriptives numériques pour résumer les données. La plus courante est la moyenne. En utilisant les données sur le PIB par tête des 60 pays figurant dans le tableau 1.1, on peut calculer la moyenne en additionnant le PIB par tête des 60 pays et en divisant la somme par 60. Le PIB par tête moyen s'élève à 21 387 dollars. Cette moyenne fournit une mesure de la valeur centrale des données.

Dans de nombreux domaines, l'intérêt pour les méthodes statistiques qui peuvent être utilisées pour développer et présenter des statistiques descriptives, continue de croître. Les chapitres 2 et 3 sont consacrés aux méthodes de statistiques descriptives sous forme de tableaux, de graphiques et sous forme numérique.

1.6 INFÉRENCE STATISTIQUE

De nombreuses situations requièrent des données relatives à un vaste ensemble d'éléments (individus, sociétés, électeurs, ménages, produits, clients, etc.). À cause de considérations telles que les coûts ou le temps, les données ne peuvent être collectées qu'auprès d'une petite partie du groupe concerné. Le groupe considéré dans son ensemble est désigné par le terme *population* et la petite partie du groupe, par le terme *échantillon*. Formellement, on utilise les définitions suivantes.

► **Population**
Une *population* est l'ensemble de tous les éléments d'intérêt dans une étude particulière.

► **Échantillon**
Un *échantillon* est un sous-ensemble de la population.

Le gouvernement américain effectue un recensement tous les dix ans. Les sociétés d'études de marché réalisent des enquêtes à partir d'échantillons de la population tous les jours.

Le processus d'enquête pour collecter des données relatives à la population entière est appelé **recensement**. Le processus d'enquête pour collecter des données relatives à un échantillon est appelé **enquête d'échantillonnage**. L'apport majeur des statistiques réside dans la possibilité de faire des estimations et des tests d'hypothèses sur les caractéristiques d'une population à partir d'un échantillon, au travers d'un processus dit d'**inférence statistique**.

Comme exemple d'inférence statistique, considérons l'étude faite par Norris Electronics. La société Norris fabrique des ampoules à forte intensité, utilisées dans de nombreux produits électriques. Dans le but d'accroître la durée de vie des ampoules, le groupe de recherche a mis au point un nouveau filament. Dans ce cas, la population

correspond à l'ensemble des ampoules produites avec le nouveau filament. Pour évaluer les performances de ce nouveau filament, 200 nouvelles ampoules ont été fabriquées et testées. Les données collectées à partir de cet échantillon indiquent le nombre d'heures d'éclairage obtenues avec chaque ampoule avant que le filament ne grille. Les données de l'échantillon sont reportées dans le tableau 1.5.

Supposons que Norris veuille utiliser les données de l'échantillon pour estimer le nombre moyen d'heures d'éclairage de toutes les ampoules qui pourraient être fabriquées avec le nouveau filament. En additionnant les 200 valeurs du tableau 1.5 et en divisant le total par 200, on obtient la durée de vie moyenne des ampoules de l'échantillon : 76 heures. La figure 1.7 résume sous forme de graphique le processus d'inférence statistique utilisé par Norris Electronics.

Quand les statisticiens utilisent un échantillon pour estimer une caractéristique de la population, ils définissent également la qualité ou précision de l'estimation. Pour l'exemple de Norris, le statisticien doit préciser que l'estimation ponctuelle de la durée de vie moyenne des ampoules de la population est de 76 heures avec une marge d'erreur de plus ou moins 4 heures. Ainsi, l'intervalle d'estimation de la durée de vie moyenne de toutes les ampoules produites est compris entre 72 et 80 heures. Le statisticien peut

Tableau 1.5 ***Nombre d'heures d'éclairage avant que l'ampoule ne grille pour un échantillon de 200 ampoules de Norris Electronics***

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73

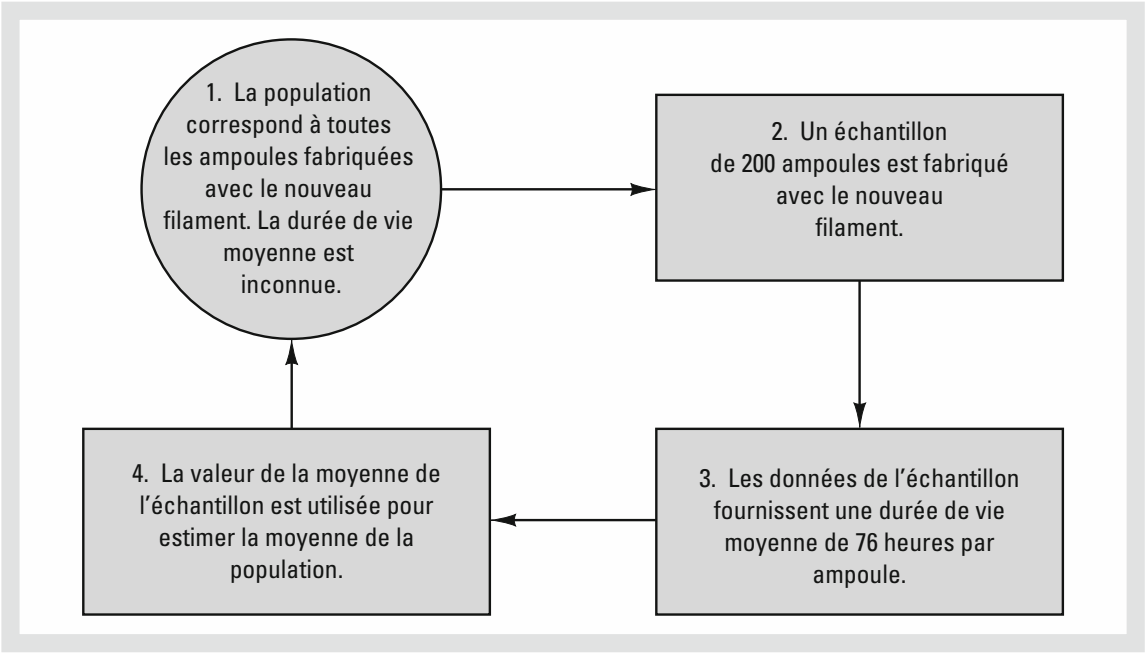


Figure 1.7 ***Le processus d'inférence statistique dans le cadre de l'exemple de Norris Electronics***

également indiquer son degré de confiance quant au fait que l'intervalle [72 ; 80] contienne la moyenne de la population.

1.7 INFORMATIQUE ET ANALYSE STATISTIQUE

Dans la mesure où l'analyse statistique implique souvent de larges ensembles de données, les analystes utilisent fréquemment des logiciels informatiques pour ce travail. Par exemple, calculer la durée de vie moyenne des 200 ampoules dans l'exemple de Norris Electronics (cf. tableau 1.5) pourrait s'avérer pénible sans un ordinateur. Pour faciliter l'usage de l'informatique, les grands ensembles de données présents dans cet ouvrage sont disponibles en ligne. Les fichiers de données sont téléchargeables à la fois au format Minitab et au format Excel. En outre, l'outil StatTools d'Excel peut être téléchargé à partir du site. Les instructions pour exécuter les procédures statistiques en utilisant Minitab, Excel et StatTools sont fournies en annexe des chapitres.

1.8 TRAITEMENT DES DONNÉES

Grâce aux lecteurs de cartes magnétiques, aux scanners des codes-barres et aux terminaux de vente, la plupart des sociétés obtiennent de nombreuses informations quotidiennes. Même pour un petit restaurant local qui utilise des tablettes tactiles pour enregistrer les commandes et délivrer l'addition, la quantité de données collectées peut être importante.

Pour les grandes enseignes de la distribution, le volume de données collectées est tel qu'il est difficile de conceptualiser comment exploiter de façon efficace ces données pour améliorer la rentabilité de l'entreprise. Par exemple, les grandes surfaces comme Walmart collectent des données relatives à 20 ou 30 millions de transactions chaque jour, les sociétés de télécommunications comme France Télécom et AT&T acheminent plus de 300 millions d'appels par jour et Visa gère 6 800 transactions de paiement par seconde, soit approximativement 600 millions de transactions par jour. Stocker et exploiter ces données est une tâche titanesque.

Le terme « stockage de données » est utilisé pour faire référence au processus de collecte, stockage et gestion des données. La puissance des ordinateurs et les outils de collecte des données ont atteint un tel niveau de développement qu'il est maintenant envisageable de stocker et de traiter des quantités très importantes de données en quelques secondes. L'analyse de données contenues dans une banque de données peut conduire à des changements de stratégie et à une augmentation des profits.

Les études relatives au traitement des données visent à développer des méthodes permettant de tirer des informations utiles à la prise décision de ces grandes bases de données. En associant des procédures statistiques, mathématiques et informatiques, les analystes exploitent les banques de données pour les convertir en informations utiles. Kurt Thearling, un pionnier dans ce domaine, définit le traitement des données comme « l'extraction automatisée d'informations prédictives à partir de grandes bases de données ». Les deux mots clés dans la définition de M. Thearling sont « automatisée » et « prédictives ». Les systèmes de traitement des données les plus efficaces utilisent des procédures automatisées pour extraire de l'information des données en utilisant seulement les requêtes, générales voire vagues, formulées par l'utilisateur. Et les logiciels de traitement des données automatisent le processus de découverte de l'information prédictive cachée, ce qui, par le passé, nécessitait des heures d'analyse.

Les applications majeures du traitement des données ont été développées par des sociétés commerciales (orientées vers les clients), telles que les commerces de détail, les organismes financiers et les opérateurs de télécommunication. Le traitement des données a été utilisé avec succès pour aider des vendeurs tels qu'Amazon et Barnes & Noble à prédire quels produits connexes les consommateurs sont susceptibles d'acheter en fonction de leurs achats passés. Grâce à cela, lorsqu'un client se connecte au site Internet d'une société et achète un produit, des fenêtres pop-up l'alertent de l'existence de produits complémentaires susceptibles de l'intéresser. Le traitement des données peut également être utilisé pour identifier les clients qui sont susceptibles de dépenser plus de 20 dollars lors d'un achat. Ces clients pourront alors bénéficier d'offres de réduction envoyées par e-mail ou par courrier, pour les inciter à renouveler leurs achats avant une certaine date.

Le traitement des données est une technologie qui repose sur des méthodes statistiques telles que les régressions multiples, les régressions logistiques et la corrélation. Il combine de façon originale toutes ces méthodes et les technologies informatiques pour optimiser le traitement des données. Un investissement significatif en temps et en argent est nécessaire pour créer des logiciels de traitement des données similaires à ceux

développés par des entreprises telles que Oracle, Teradata et SAS. Les concepts statistiques introduits dans cet ouvrage vous seront utiles pour comprendre la méthodologie statistique utilisée par les logiciels de traitement des données et vous permettront de mieux comprendre l'information statistique qui est fournie.

Les méthodes statistiques jouent un rôle important dans le traitement des données, à la fois en termes de découverte des relations entre les données et de prédiction des résultats futurs. Cependant, une étude approfondie des techniques et méthodes de traitement des données est hors du champ de cet ouvrage.

Dans la mesure où les modèles statistiques jouent un rôle important dans le développement des modèles prédictifs, les statisticiens doivent prendre un certain nombre de précautions pour correctement formuler ces modèles statistiques. Par exemple, la question de la fiabilité du modèle est une question primordiale. Un modèle statistique qui fonctionne bien pour un échantillon particulier de données ne pourra pas nécessairement être appliqué de façon fiable à d'autres jeux de données. Une des approches statistiques courantes pour évaluer la fiabilité d'un modèle consiste à diviser l'ensemble des données d'échantillon en deux sous-ensembles : un sous-ensemble de données d'entraînement et un sous-ensemble de données de test. Si le modèle développé en utilisant les données d'entraînement est capable de prédire avec précision les données de test, on dit que le modèle est fiable. Un avantage qu'a le traitement des données par rapport aux statistiques classiques, réside dans la quantité astronomique de données disponibles. Cela permet au logiciel de traitement des données de séparer l'ensemble des données de façon à tester la fiabilité d'un modèle développé sur un sous-ensemble de données d'entraînement sur d'autres données. En ce sens, la séparation de l'ensemble des données en plusieurs sous-ensembles permet de développer des modèles, d'établir des relations entre les variables et ensuite d'observer rapidement si ces modèles et relations sont reproductibles et valables avec des données différentes. Le risque en ayant tant de données réside dans la détermination d'association et de relation de cause à effet qui n'existent pas réellement. Une interprétation prudente des résultats obtenus via les procédures de traitement des données et des tests supplémentaires aideront à éviter cet écueil.

1.9 GUIDE DES BONNES PRATIQUES STATISTIQUES

On doit s'efforcer d'avoir un comportement éthique exemplaire dans tout ce que l'on fait. Des questions éthiques surgissent en statistiques à cause du rôle important des statistiques dans la collecte, l'analyse, la présentation et l'interprétation des données. Dans une étude statistique, des comportements non-éthiques peuvent prendre différentes formes : échantillonnage inapproprié, analyse biaisée des données, développement de graphiques trompeurs, utilisation de statistiques descriptives inappropriées et/ou interprétation biaisée des résultats statistiques.

Nous vous encourageons, dans votre propre travail statistique, à être équitable, minutieux, objectif et neutre, à la fois lorsque vous collectez des données, effectuez des

analyses, faîtes des présentations orales et rédigez des rapports. En tant que consommateur de statistiques, vous devez également être conscient de la possibilité que certains statisticiens n'aient pas un comportement éthique. Lorsque vous êtes confrontés à des statistiques dans les journaux, à la télévision, sur Internet, etc., il est judicieux d'avoir un certain recul sur ces informations, de toujours tenir compte des sources, du but et de l'objectivité des statistiques fournies.

L'association américaine de statistiques, la principale organisation statistique professionnelle des États-Unis, a rédigé un rapport intitulé *Ethical Guidelines for Statistical Practice*[2]. Ce guide a vocation à aider les statisticiens à travailler de façon éthique et responsable. Le rapport contient 67 recommandations organisées en huit items : professionnalisme ; responsabilités vis-à-vis des commanditaires, clients et employeurs ; responsabilités lors des publications et témoignages ; responsabilités vis-à-vis des sujets de recherche ; responsabilités vis-à-vis de l'équipe de recherche ; responsabilité vis-à-vis des autres statisticiens ; responsabilités relatives aux allégations de mauvaises conduites ; et responsabilités des organisations, des individus, des avocats et autres clients qui emploient des statisticiens.

L'une des recommandations éthiques dans le domaine du professionnalisme soulève la question de la conduite de tests multiples jusqu'à ce que le résultat désiré soit obtenu. Considérons un exemple. Dans la section 1.5, nous avons évoqué un test statistique effectué par Norris Electronics impliquant un échantillon de 200 ampoules à haute intensité fabriquées avec un nouveau filament. La durée de vie moyenne de l'échantillon, 76 heures, fournit une estimation de la durée de vie moyenne de toutes les ampoules fabriquées avec le nouveau filament. Cependant, puisque Norris a sélectionné un échantillon d'ampoules, il est raisonnable de supposer qu'un autre échantillon aurait fourni une durée de vie moyenne différente.

Supposez que la direction de Norris ait espéré que les résultats de l'échantillon lui permettraient de déclarer que la durée de vie moyenne des nouvelles ampoules est d'au moins 80 heures. Supposez par ailleurs que la direction de Norris décide de poursuivre l'étude en fabriquant et en testant des échantillons différents de 200 ampoules fabriquées avec le nouveau filament jusqu'à ce qu'une moyenne d'échantillon d'au moins 80 heures soit obtenue. Si l'étude est répétée un nombre suffisant de fois, un échantillon peut éventuellement – uniquement par chance – fournir le résultat désiré et permettre à Norris de faire une telle déclaration. Dans ce cas, les clients pourraient être amenés à croire (de façon erronée) que le nouveau produit est meilleur que le produit actuel. Clairement, ce type de comportement est non-éthique et représente une mauvaise utilisation des statistiques en pratique.

Plusieurs recommandations éthiques dans le domaine des responsabilités et des publications traitent de questions relatives au traitement des données. Par exemple, un statisticien doit tenir compte de toutes les données considérées dans une étude et décrire le (ou les) échantillon(s) réellement utilisé(s). Dans l'étude de Norris Electronics, la durée de vie moyenne pour les 200 ampoules dans l'échantillon originel est de 76 heures ; c'est considérablement moins que les 80 heures ou plus que la direction espérait atteindre. Supposez maintenant qu'après avoir revu les résultats établissant une durée de vie moyenne de

2 Association américaine de statistiques, *Ethical Guidelines for Statistical Practice*, 1999.

76 heures, Norris écarte toutes les observations inférieures ou égales à 70 heures (avant que l'ampoule ne grille), en décrétant que ces ampoules contiennent des imperfections liées à la phase de démarrage du processus de fabrication. Après avoir écarté ces ampoules, la durée de vie moyenne des ampoules restantes dans l'échantillon s'élève à 82 heures. Douteriez-vous d'une déclaration de Norris affirmant que la durée de vie moyenne de ses ampoules est de 82 heures ?

Si les ampoules de Norris dont la durée de vie est inférieure ou égale à 70 heures sont écartées dans le but de fournir une durée de vie moyenne de 82 heures, cette mise à l'écart de certaines observations est incontestablement contraire à l'éthique. Mais, même si les ampoules écartées contiennent des imperfections générées par des problèmes survenus au démarrage du processus de fabrication – et, par conséquent, ne devraient pas être incluses dans l'analyse – le statisticien qui effectue l'étude doit tenir compte de toutes les données observées et expliquer comment l'échantillon utilisé a été obtenu. Avoir une autre démarche est potentiellement dangereux et peut constituer un comportement non-éthique de la part à la fois de la société et du statisticien.

Une des recommandations du rapport de l'association américaine de statistiques stipule que les statisticiens doivent éviter toute tendance à orienter le travail statistique vers des résultats prédéterminés. Ce type de pratique non éthique est souvent observé lorsque des échantillons non représentatifs sont utilisés pour établir des affirmations. Par exemple, dans de nombreux États américains, fumer dans les restaurants est interdit. Supposez qu'un lobbyiste de l'industrie du tabac interroge des personnes dans les restaurants où fumer est autorisé, dans le but d'estimer le pourcentage de personnes en faveur du tabac dans les restaurants. Les résultats de l'échantillon montrent que 90 % des personnes interrogées sont favorables au tabac dans les restaurants. En se basant sur les résultats de cet échantillon, le lobbyiste affirme que 90 % des personnes qui fréquentent des restaurants sont favorables au tabac dans les restaurants. Dans ce cas, on peut rétorquer que n'échantillonner que les personnes fréquentant des restaurants où fumer est autorisé, biaise les résultats. Si seuls les résultats d'une telle étude sont rapportés, les lecteurs qui ne connaissent pas les détails de l'étude (c'est-à-dire que l'échantillon n'a été collecté que dans les restaurants autorisant de fumer) peuvent être abusés.

Le contenu du rapport de l'association américaine de statistiques est large et inclut des recommandations éthiques qui sont appropriées non seulement pour un statisticien mais aussi pour les consommateurs de statistiques. Nous vous encourageons à lire ce rapport pour mieux appréhender les questions d'éthique et mettre en application ces principes éthiques lorsque vous ferez vos propres analyses.

RÉSUMÉ

Les statistiques sont l'art et la science de collecter, analyser, présenter et interpréter des données. Pratiquement tous les étudiants en économie ou en commerce suivent des cours de statistique. Nous avons débuté ce chapitre par une présentation des applications statistiques usuelles en économie et dans le domaine commercial.

Les données sont les faits et les chiffres qui sont collectés et analysés. Il existe quatre échelles de mesure utilisées pour obtenir des données sur une variable particulière : nominale, ordinale, cardinale (par intervalle) ou de rapport. L'échelle de mesure d'une variable est nominale lorsque des labels ou des noms permettent d'identifier une caractéristique d'un élément. L'échelle est ordinale si les données ont les propriétés nominales et si l'ordre ou le rang des données fait sens. L'échelle est dite cardinale (par intervalle) si les données possèdent les propriétés ordinales et si l'intervalle entre les valeurs est mesuré selon une unité fixe. Enfin, l'échelle de mesure est dite de rapport si les données possèdent les propriétés de données cardinales et si le rapport entre deux valeurs est porteur de sens.

Dans une perspective d'analyse, les données peuvent être classées selon leur nature quantitative ou qualitative. Les données qualitatives utilisent des étiquettes ou des noms pour identifier une caractéristique de chaque élément. Les données qualitatives ont une échelle de mesure nominale ou ordinale et peuvent être numériques ou non numériques. Les données quantitatives sont des valeurs numériques qui indiquent des quantités. Les données quantitatives sont évaluées grâce à une échelle de mesure cardinale (par intervalle) ou de rapport. Les opérations arithmétiques ordinaires ne sont pertinentes qu'avec des variables quantitatives. Ainsi, les opérations statistiques utilisées pour des données quantitatives ne sont pas toujours appropriées pour des données qualitatives.

Dans les sections 1.4 et 1.5, nous avons abordé les sujets de statistique descriptive et d'inférence statistique. Les statistiques descriptives sont constituées de tableaux, de graphiques ou de chiffres résumant les données. L'inférence statistique est le processus qui consiste à utiliser les données d'un échantillon pour effectuer des estimations ou des tests d'hypothèses concernant les caractéristiques d'une population. Les trois dernières sections de ce chapitre fournissent des informations sur le rôle des ordinateurs dans l'analyse statistique, une introduction à la discipline relativement récente de traitement des données et un résumé des recommandations éthiques pour la pratique des statistiques.

GLOSSAIRE

Statistiques. L'art et la science de collecter, analyser, présenter et interpréter des données.

Données. Faits et chiffres qui sont collectés, analysés et résumés pour être présentés et interprétés.

Ensemble de données. Toutes les données collectées pour une étude particulière.

Éléments. Entités sur lesquelles portent la collecte de données.

Variable. Caractéristique des éléments qui nous intéresse.

Observation. Ensemble des mesures obtenues pour un élément unique.

Échelle nominale. Échelle de mesure d'une variable dont les données sont des labels ou noms identifiant une caractéristique d'un élément. Les données nominales peuvent être numériques ou non.

Échelle ordinale. Échelle de mesure d'une variable dont les données possèdent les propriétés nominales et dont l'ordre fait sens. Les données ordinales peuvent être numériques ou non.

ÉCHELLE CARDINALE OU D'INTERVALLE. Échelle de mesure d'une variable dont les données possèdent les propriétés ordinales et dont l'écart peut être exprimé selon une unité de mesure fixe. Les données cardinales sont toujours numériques.

ÉCHELLE DE RAPPORT. Échelle de mesure d'une variable dont les données possèdent les propriétés cardinales et dont le rapport fait sens. Les données mesurées selon une échelle de rapport sont toujours numériques.

DONNÉES QUALITATIVES (OU CATÉGORIELLES). Labels ou noms utilisés pour identifier une caractéristique de chaque élément de l'ensemble de données. Les données qualitatives utilisent une échelle de mesure nominale ou ordinale et peuvent être numériques ou non numériques.

DONNÉES QUANTITATIVES. Valeurs numériques qui indiquent la quantité de quelque chose. Les données quantitatives sont mesurées selon une échelle cardinale ou de rapport.

VARIABLE QUALITATIVE (OU CATÉGORIELLE). Variable dont les données sont qualitatives.

VARIABLE QUANTITATIVE. Variable dont les données sont quantitatives.

DONNÉES EN COUPE TRANSVERSALE. Données collectées à un même moment (ou à des moments très proches) dans le temps.

DONNÉES DE SÉRIE TEMPORELLE. Données collectées à des moments différents dans le temps.

STATISTIQUES DESCRIPTIVES. Tableaux, graphiques et approches numériques utilisés pour résumer les données.

POPULATION. Ensemble de tous les éléments d'intérêt dans une étude particulière.

ÉCHANTILLON. Sous-ensemble de la population.

RECENSEMENT. Enquête visant à collecter des données relatives à la population entière.

ENQUÊTE D'ÉCHANTILLONNAGE. Enquête visant à collecter des données relatives à un échantillon.

INFÉRENCE STATISTIQUE. Processus d'utilisation des données d'un échantillon pour estimer ou tester des hypothèses sur les caractéristiques d'une population.

TRAITEMENT DES DONNÉES. Processus d'utilisation de procédures issues des statistiques et de l'informatique pour extraire des informations utiles de bases de données très importantes.

EXERCICES

1. Discuter des différences entre les statistiques en tant que faits numériques et les statistiques en tant que discipline ou objet d'étude.

2. Le département américain à l'énergie fournit des informations sur le prix des carburants pour différents types de moteurs. Un échantillon de 10 automobiles est fourni dans le tableau 1.6 (site Internet de Fuel Economy, 22 février 2008). Les données indiquent la taille du véhicule (compacte, moyenne ou grande), la puissance du moteur (nombre de chevaux), la consommation en ville (nombre de miles parcourus avec un gallon de carburant), la consommation sur autoroute (nombre de miles parcourus avec un gallon de carburant) et le type de carburant recommandé (diesel, sans plomb ou ordinaire).

a) Combien d'éléments y a-t-il dans l'ensemble de données ?

b) Combien de variables y a-t-il dans l'ensemble de données ?

c) Quelles sont les variables qualitatives ? Quelles sont les variables quantitatives ?

d) Quel type d'échelle de mesure est utilisé pour chacune des variables ?

Tableau 1.6 ***Information sur la consommation de carburant de 10 véhicules***

Marque	Taille	Chevaux	Consommation urbaine	Consommation sur autoroute	Carburant
Audi A8	Grande	12	13	19	Sans plomb
BMW 328Xi	Compacte	6	17	25	Sans plomb
Cadillac CTS	Moyenne	6	16	25	Ordinaire
Chrysler 300	Grande	8	13	18	Sans plomb
Ford Focus	Compacte	4	24	33	Ordinaire
Hyundai Elantra	Moyenne	4	25	33	Ordinaire
Jeep Grand Cherokee	Moyenne	6	17	26	Diesel
Pontiac G6	Compacte	6	15	22	Ordinaire
Toyota Camry	Moyenne	4	21	31	Ordinaire
Volkswagen Jetta	Compacte	5	21	29	Ordinaire

3. Reprendre les données du tableau 1.6.

a) Quelle est la consommation moyenne en ville ?

b) En moyenne, quel est l'écart de consommation en zone urbaine et sur autoroute ?

c) Quel est le pourcentage de voitures ayant des moteurs de 4 chevaux ?

d) Quel est le pourcentage de voitures utilisant du carburant ordinaire ?

4. Le tableau 1.7 fournit des données relatives à huit téléphones sans fil (*Consumer Reports*, novembre 2012). La note globale, une mesure de la qualité globale du téléphone sans fil, varie entre 0 et 100. La qualité sonore peut être mauvaise, satisfaisante, bonne, très bonne ou excellente. L'autonomie correspond au nombre d'heures durant lesquelles le téléphone peut être utilisé, lorsqu'il est complètement chargé, selon les dires du fabricant.

Tableau 1.7 ***Données relatives à huit téléphones sans fil***

Marque	Modèle	Prix (dollars)	Note globale	Qualité sonore	Combiné sur base	Autonomie (heures)
AT&T	CL84100	60	73	Excellente	Oui	7
AT&T	TL92271	80	70	Très bonne	Non	7
Panasonic	4773B	100	78	Très bonne	Oui	13
Panasonic	6592T	70	72	Très bonne	Non	13
Uniden	D2997	45	70	Très bonne	Non	10
Uniden	D1788	80	73	Très bonne	Oui	7
Vtech	DS6521	60	72	Excellente	Non	7
Vtech	CS6649	50	72	Très bonne	Oui	7

a) Combien d'éléments y a-t-il dans cet ensemble de données ?

b) Parmi les variables Prix, Note globale, Qualité sonore, Combiné sur base et Autonomie, lesquelles sont quantitatives, lesquelles sont qualitatives ?

c) Quelle est l'échelle de mesure utilisée pour chacune de ces variables ?

5. Reprendre l'ensemble de données du tableau 1.7.

a) Quel est le prix moyen de ces téléphones sans fil ?

b) Quelle est l'autonomie moyenne de ces téléphones sans fil ?

c) Quel est le pourcentage de téléphones sans fil qui ont une excellente qualité sonore ?

d) Quel est le pourcentage de téléphones sans fil qui ont un combiné sur base ?

6. J.D. Power et Associés effectue des sondages auprès des propriétaires d'une nouvelle voiture pour déterminer la qualité de leur véhicule récemment acheté. Les questions suivantes ont été posées dans l'enquête *J.D. Power Initial Quality Survey*, réalisée en mai 2012 :

a) Avez-vous achetez ou louez-vous le véhicule ?

b) Quel prix avez-vous payé ?

c) Comment qualifieriez-vous l'apparence extérieure de votre voiture ? (Moche, Moyenne, Exceptionnelle ou Vraiment exceptionnelle)

d) Quelle est sa consommation moyenne (nombre de miles parcourus avec un gallon de carburant) ?

e) Quelle note globale donneriez-vous à votre nouvelle voiture ? (entre 1 et 10 points, 1 pour insuffisante et 10 pour vraiment exceptionnelle)

Dire si chaque question fournit des données quantitatives ou qualitatives.

7. La société Kroger est l'une des plus grandes enseignes de la distribution aux États-Unis, avec plus de 2 000 magasins à travers le pays. Kroger réalise un sondage d'opinion en ligne auprès de ses clients pour obtenir des données de performance sur ses produits et services et connaître les motivations de ses clients (site Internet de Kroger, avril 2012). Dans cette enquête, on demande aux clients de Kroger s'ils seraient prêts à payer davantage pour des produits qui auraient chacune des quatre caractéristiques suivantes. Les quatre questions étaient : Seriez-vous prêts à payer davantage pour des produits de marque ? des produits qui respectent l'environnement ? des produits bio ? des produits qui vous sont recommandés par d'autres personnes ?

À chaque question, les clients pouvaient répondre Oui s'ils étaient prêts à payer davantage ou Non s'ils n'étaient pas disposés à payer plus.

a) Les données collectées par Kroger dans cet exemple sont-elles qualitatives ou quantitatives ?

b) Quelle est l'échelle de mesure utilisée ?

8. L'enquête *Financial Times*/Harris est une enquête mensuelle en ligne réalisée auprès d'adultes de six pays européens et aux États-Unis. L'enquête menée en janvier a été réalisée auprès de 1 015 adultes vivant aux États-Unis. Une des questions posées était : « Comment évalueriez-vous la Banque Fédérale dans sa gestion des problèmes de crédit sur les marchés financiers ? » Les réponses possibles étaient : excellente, bonne, correcte, mauvaise, terrible (site Internet de Harris Interactive, janvier 2008).

a) Quelle était la taille de l'échantillon de cette enquête ?

b) Les données sont-elles qualitatives ou quantitatives ?

c) Est-il plus pertinent d'utiliser des moyennes ou des pourcentages pour résumer les réponses à la question posée ?

d) Parmi les personnes ayant répondu, 10 % ont déclaré que la Banque Fédérale faisait du bon travail. Combien d'individus ont fourni cette réponse ?

9. Le département au commerce a rapporté que, parmi les prétendants au prix national de la qualité Malcolm Baldrige, 23 étaient de grandes entreprises manufacturières, 18 de grandes entreprises prestataires de service et 30 étaient de petites entreprises.

a) Le type d'entreprises est-il une variable qualitative ou quantitative ?

b) Quel est le pourcentage de candidatures émanant de petites entreprises ?

10. L'enquête auprès des ménages menée par le bureau des statistiques du transport est actualisée chaque année et constitue une source d'information pour le département américain des transports. Dans une des parties de l'enquête, on demande aux personnes interrogées de réagir à l'affirmation suivante : « Les conducteurs de véhicules motorisés devraient être autorisés à téléphoner en utilisant des kits mains-libres lorsqu'ils conduisent. » Les réponses possibles étaient : tout à fait d'accord, d'accord, pas d'accord, tout à fait pas d'accord. Quarante-quatre personnes ont répondu être tout à fait d'accord avec cette affirmation, 130 d'accord, 165 pas d'accord et 741 tout à fait pas d'accord (site Internet du bureau des transports, août 2010).

a) Les réponses à cette affirmation constituent-elles des données quantitatives ou qualitatives ?

b) Serait-il plus pertinent d'utiliser des moyennes ou des pourcentages pour résumer les réponses obtenues ?

c) Quel est le pourcentage de personnes interrogées qui sont tout à fait d'accord avec le fait d'autoriser les conducteurs de véhicules motorisés à utiliser le kit mains-libres pour téléphoner en conduisant ?

d) Les résultats indiquent-ils une tendance favorable ou défavorable à l'idée d'autoriser l'usage du téléphone avec kit mains-libres en conduisant ?

11. La société J.D. Power et associés mène des enquêtes de qualité sur les véhicules afin de fournir aux fabricants automobiles des informations sur la satisfaction des clients quant à leurs produits (*Enquête sur la qualité des véhicules*, janvier 2010). En utilisant un échantillon de propriétaires de véhicules collecté à partir des registres d'achats récents, l'enquête posait une série de questions aux propriétaires, relatives à leur nouveau véhicule telles que celles qui suivent. Pour chaque question, dire si les données collectées sont qualitatives ou quantitatives et indiquer l'échelle de mesure utilisée.

a) Quel prix avez-vous payé pour acheter votre véhicule ?

b) Comment avez-vous payé votre véhicule ? (en espèce, en location ou à crédit)

c) Recommanderiez-vous ce véhicule à un ami ? (absolument pas, probablement pas, probablement, absolument)

d) Quel est le kilométrage actuel de votre véhicule ?

e) Comment noteriez-vous globalement votre nouveau véhicule ? Une échelle de 10 points (de 1, médiocre à 10, exceptionnelle) était utilisée.

12. L'office du tourisme de Hawaii a collecté des données sur les touristes de l'île. Les questions suivantes sont extraites d'un questionnaire comportant 16 questions, distribué aux passagers d'un vol à destination de Hawaii.

- Ce voyage à Hawaii est mon : 1^{er}, 2^{e}, 3^{e}, 4^{e}, etc.
- La raison principale de ce voyage est : (10 catégories dont vacances, convention, lune de miel)
- Où est-ce que j'envisage de séjourner (11 catégories dont hôtel, appartement, dépendances, camping)
- Nombre de jours à passer à Hawaii

a) Quelle est la population étudiée ?

b) Est-ce que le questionnaire est un bon moyen d'atteindre la population des passagers d'un vol à destination d'Hawaii ?

c) Dire si chacune des quatre questions précédentes fournit des données qualitatives ou quantitatives ?

13. Le graphique 1.8 est un diagramme en barres résumant les dépenses fédérales des années 2004 à 2010 (site Internet du département du budget du Congrès, 15 mai 2011).

a) Quelle est la variable à laquelle on s'intéresse ?

b) Les données sont-elles qualitatives ou quantitatives ?

c) Les données sont-elles des données en coupe transversale ou des données de série temporelle ?

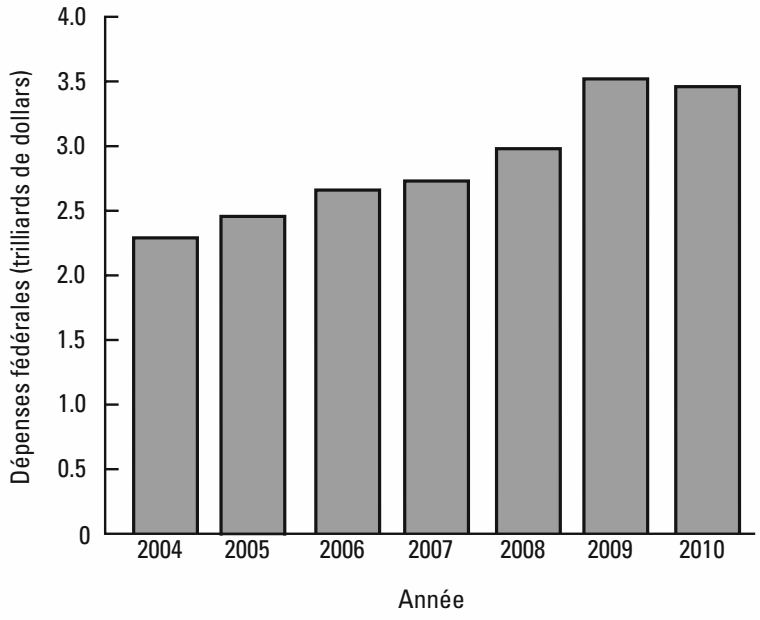


Figure 1.8 ***Dépenses fédérales***

d) Commenter l'évolution des dépenses fédérales sur la période.

14. Les données suivantes indiquent le nombre de véhicules de location en service pour trois sociétés de location de voitures : Hertz, Avis et Dollar. Les données couvrent la période 2007-2010 et sont exprimées en milliers de véhicules (site Internet de Auto Rental News, 15 mai 2011).

	Nombre de véhicules en service			
Société	**2007**	**2008**	**2009**	**2010**
Hertz	327	311	286	290
Dollar	167	140	106	108
Avis	204	220	300	270

a) Construire un graphique indiquant le nombre de voitures de location en service pour chaque société entre 2007 et 2010. Représenter ces séries temporelles pour les trois sociétés sur un même graphique.

b) Quelle est la société qui apparaît comme le leader en part de marché ? Comment les parts de marché ont-elles évolué au cours de la période ?

c) Construire un diagramme en barres représentant les voitures de location en service en 2010. Ce graphique est-il construit à partir de données en coupe transversale ou d'une série temporelle ?

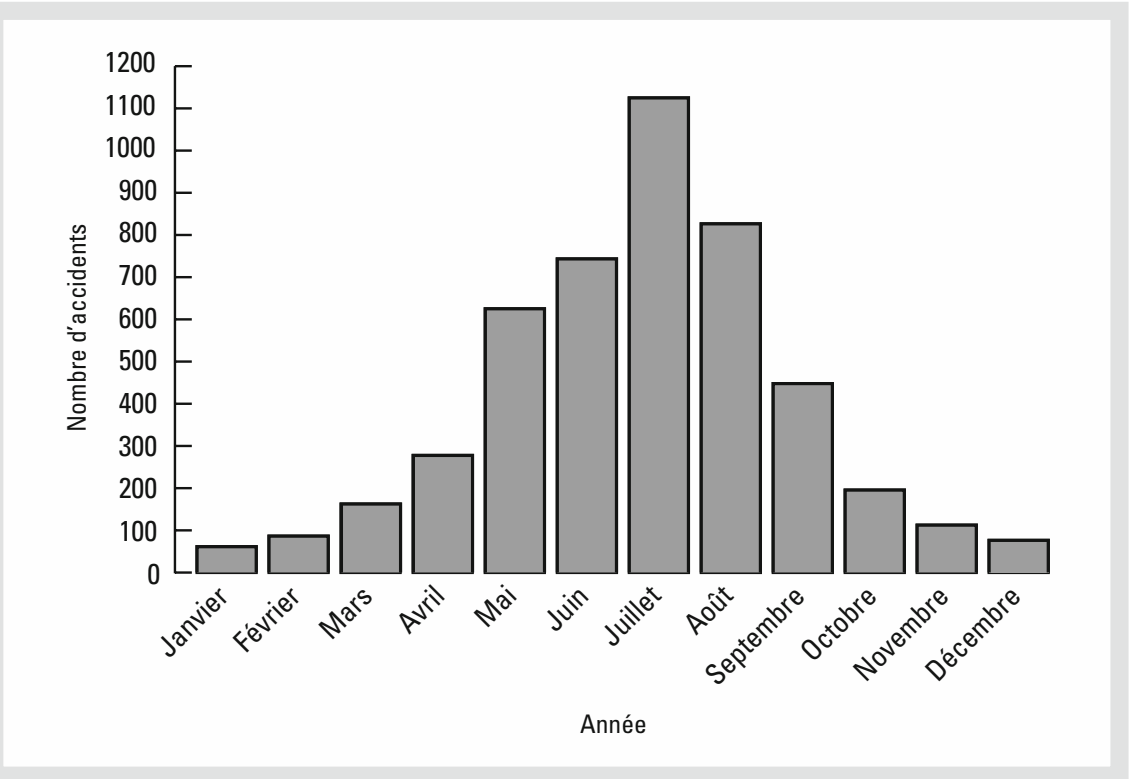


Figure 1.9 ***Nombre d'accidents impliquant des bateaux de plaisance***

15. Chaque année, les gardes côtes américains collectent des données et établissent des statistiques sur les accidents impliquant des bateaux de plaisance. Ces statistiques sont issues des rapports d'accidents rédigés par les propriétaires ou les conducteurs des bateaux de plaisance impliqués dans des accidents. En 2009, 4 730 rapports d'accidents impliquant des bateaux de plaisance ont été enregistrés. Un diagramme en barres résumant le nombre de rapports d'accidents enregistrés chaque mois est représenté à la figure 1.9 (site Internet de la division sécurité des bateaux des gardes côtes américains, août 2010).

- **a)** Les données sont-elles qualitatives ou quantitatives ?
- **b)** Les données sont-elles des données en coupe transversale ou des données de série temporelle ?
- **c)** Au cours de quel mois le plus de rapports d'accidents ont-ils été enregistrés ? Combien approximativement ?
- **d)** Soixante-et-un rapports d'accidents ont été enregistrés en janvier et 76 en décembre. Quel pourcentage du nombre total d'accidents enregistrés au cours de l'année a été enregistré au cours de ces deux mois ? Ce résultat vous semble-t-il raisonnable ?
- **e)** Commenter la forme générale du graphique.

16. Le service d'information sur l'énergie du Département américain de l'énergie fournissait des séries temporelles du prix moyen d'un gallon d'essence sans plomb entre janvier 2007 et mars 2012 (site Internet du service d'information sur l'énergie, avril 2012). Utilisez Internet pour obtenir le prix moyen d'un gallon d'essence sans plomb depuis mars 2012.

- **a)** Poursuivez le graphique présenté à la figure 1.1.
- **b)** Quelles interprétations pouvez-vous faire du prix moyen par gallon de l'essence sans plomb depuis mars 2012 ?
- **c)** Les données indiquent-elles une poursuite de l'augmentation des prix durant les mois d'été ? Expliquez.

17. Le manager d'une grande entreprise a recommandé d'augmenter le salaire d'un employé de grande valeur de 10 000 dollars pour le dissuader de quitter l'entreprise. Quelles sources de données internes et externes devraient être utilisées pour décider si une telle augmentation de salaire est appropriée ?

18. Un sondage aléatoire mené par téléphone auprès de 1 021 adultes (âgés de 18 ans et plus) a été effectué par Opinion Research Corporation pour le compte de CompleteTax, un service en ligne d'aide pour effectuer sa déclaration d'impôt. Les résultats du sondage indiquent que 684 des personnes interrogées envisageaient d'effectuer leur déclaration d'impôt électroniquement (enquête CompleteTax de 2010).

- **a)** Développer une statistique descriptive qui permet d'estimer le pourcentage de contribuables qui effectuent leur déclaration par Internet.
- **b)** L'enquête rapporte que le moyen le plus fréquemment utilisé par les contribuables pour les aider à préparer leur déclaration est le recours aux services d'un comptable ou d'un fiscaliste. Si 60 % des personnes interrogées préparent leur déclaration de cette façon, combien ont eu recours à un comptable ou un fiscaliste ?
- **c)** Les autres méthodes pour aider une personne à faire sa déclaration incluent une préparation manuelle, l'utilisation d'un service fiscal en ligne et l'utilisation d'un

logiciel informatique de taxation. Les données sur les méthodes de préparation au remplissage des déclarations sont-elles quantitatives ou qualitatives ?

19. L'enquête réalisée auprès des abonnés Nord-Américains par *Bloomberg Businessweek* a permis de collecter des données sur un échantillon de 2 861 abonnés. Cinquante-neuf pourcent des personnes ayant répondu à l'enquête ont indiqué que leur salaire annuel était supérieur à 75 000 $ et plus de 50 % ont déclaré posséder une carte de crédit American Express.

a) Quelle est la population concernée dans cette étude ?

b) Est-ce que le revenu annuel est une variable qualitative ou quantitative ?

c) Est-ce que la possession d'une carte de crédit American Express est une variable qualitative ou quantitative ?

d) Est-ce que les données de cette étude sont en coupe transversale ou sont des séries temporelles ?

e) Décrire quelques inférences statistiques que *Bloomberg Businessweek* pourrait faire sur la base de cette étude.

20. Une enquête réalisée auprès de 131 investisseurs dans le cadre du sondage Big Money de *Barron's* révélait que :

- 43 % des investisseurs considéraient la tendance sur le marché boursier comme étant haussière ou très haussière.
- Le rendement moyen attendu des actions sur les douze mois suivants était de 11,2 %.
- 21 % des investisseurs considéraient le secteur médical comme celui qui tirerait le marché au cours des douze mois suivants.
- Lorsque l'on demandait aux investisseurs combien de temps les titres des secteurs technologiques et des télécommunications mettraient pour retrouver une croissance soutenable, leur réponse moyenne était deux ans et demi.

a) Citer deux statistiques descriptives.

b) Inférer le rendement moyen des actions attendu par la population de tous les investisseurs au cours des douze mois suivants.

c) Inférer la durée moyenne qu'il faudra aux titres technologiques et de télécommunications pour retrouver une croissance soutenable.

21. Une étude médicale de sept ans a conclu que les femmes dont les mères consommaient de la drogue DES au cours de leur grossesse étaient deux fois plus à même de développer des anormalités au niveau des tissus pouvant provoquer un cancer, que les femmes dont les mères ne prenaient pas cette drogue.

a) Cette étude implique la comparaison de deux populations. Quelles sont ces populations ?

b) Pensez-vous que les données ont été obtenues par une étude ou une expérimentation ?

c) Parmi la population des femmes dont les mères prenaient la drogue DES au cours de leur grossesse, sur un échantillon de 3 980 femmes, 63 avaient développé des anormalités au niveau des tissus qui pouvaient provoquer un cancer. Fournir une statistique descriptive qui peut servir à estimer le nombre de femmes sur 1 000 dans cette population qui ont des anormalités au niveau des tissus.

d) Pour la population des femmes dont les mères ne prenaient pas la drogue DES au cours de leur grossesse, quelle est l'estimation du nombre de femmes sur 1 000 qui pourraient avoir développé des anormalités au niveau des tissus ?

e) Les études médicales utilisent souvent un échantillon relativement grand (dans ce cas, 3 980). Pourquoi ?

22. Le centre de recherche Pew est un institut de sondage indépendant qui fournit des informations sur les problématiques, les attitudes et les tendances qui modèlent l'Amérique. Dans une enquête récente, 47 % des adultes américains ont déclaré lire une partie des informations locales sur leur téléphone ou leur tablette (site Internet de Pew, 14 mai 2011). De plus, 42 % des personnes interrogées qui possèdent un téléphone ou une tablette ont déclaré utiliser ces appareils pour s'informer de la météo locale et 37 % pour trouver un restaurant ou d'autres commerces dans les environs.

a) Une des statistiques concernait l'utilisation des téléphones ou des tablettes pour prendre connaissance des informations locales. À quelle population s'applique cette statistique ?

b) Une autre statistique concernait l'utilisation des téléphones ou des tablettes pour s'informer de la météo locale et trouver des restaurants à proximité. À quelle population s'applique cette statistique ?

c) Pensez-vous que les chercheurs de Pew ont effectué un recensement ou un sondage auprès d'un échantillon pour obtenir ces résultats ? Pourquoi ?

d) Si vous êtes propriétaire d'un restaurant, trouveriez-vous ces résultats intéressants ? Pourquoi ? Comment pourriez-vous exploiter ces informations ?

23. Nielsen Media Research mène chaque semaine des enquêtes sur l'audimat télévisuel à travers les États-Unis et publie à la fois les taux d'audience et les parts de marché. Le taux d'audience de Nielsen correspond au pourcentage de ménages possédant une télévision qui regardent un programme défini, alors que la part de marché correspond au pourcentage de ménages regardant un programme particulier parmi l'ensemble des ménages regardant la télévision. Par exemple, lors du match de baseball entre les New York Yankees et les Florida Marlins en 2003, le taux d'audience fut de 12,8 % et la part de marché de 22 % (*Associated Press*, 27 octobre 2003). Ainsi, 12,8 % des ménages possédant une télévision ont regardé le match et 22 % des ménages regardant la télévision regardaient précisément le match. En se basant sur les taux d'audience et les parts de marché des principaux programmes de télévision, Nielsen publie chaque semaine un classement des programmes ainsi qu'un classement des quatre plus grandes chaînes : ABC, CBS, NBC et Fox.

a) Qu'est-ce que la société Nielsen essaie de mesurer ?

b) Quelle est la population ?

c) Pourquoi est-il nécessaire d'utiliser un échantillon dans cette étude ?

d) Quelles sortes de décisions ou d'actions sont basées sur les études Nielsen ?

24. Un échantillon des notes obtenues lors de l'examen trimestriel de cinq étudiants fournit les données suivantes : 72, 65, 82, 90, 76. Parmi les affirmations suivantes, lesquelles sont correctes et lesquelles peuvent être qualifiées de trop générale ?

a) La moyenne des notes obtenues par l'échantillon des cinq étudiants est de 77.

b) La moyenne des notes de tous les étudiants qui ont passé leur examen est de 77.

c) Une estimation de la moyenne des notes de tous les étudiants qui ont passé leur examen est de 77.

d) Plus de la moitié des étudiants qui ont passé leur examen ont des notes comprises entre 70 et 85.

e) Si cinq autres étudiants étaient inclus dans l'échantillon, leurs notes seraient comprises entre 65 et 90.

25. Le tableau 1.8 contient un ensemble de données fournissant des informations sur 25 titres du marché secondaire listés par l'Association américaine des investisseurs individuels. Les titres du marché secondaire sont souvent des titres de sociétés plus petites qui ne sont

Tableau 1.8 ***Données pour un ensemble de 25 titres secondaires***

Société	Place boursière	Symbole	Capitalisation boursière (millions de dollars)	Coefficient de capitalisation des résultats	Marge brute (%)
DeWolfe Companies	AMEX	DWL	36,4	8,4	36,7
North Coast Energy	OTC	NCEB	52,5	6,2	59,3
Hansen Natural Corp.	OTC	HANS	41,1	14,6	44,8
MarineMax, Inc.	NYSE	HZO	111,5	7,2	23,8
Nanometrics Incorporated	OTC	NANO	228,6	38,0	53,3
TeamStaff, Inc.	OTC	TSTF	92,1	33,5	4,1
Environmental Tectonics	AMEX	ETC	51,1	35,8	35,9
Measurement Specialties	AMEX	MSS	101,8	26,8	37,6
SEMCO Energy, Inc.	NYSE	SEN	193,4	18,7	23,6
Party City Corporation	OTC	PCTY	97,2	15,9	36,4
Embrex, Inc.	OTC	EMBX	136,5	18,9	59,5
Tech/Ops Sevcon, Inc.	AMEX	TO	23,2	20,7	35,7
ARCADIS NV	OTC	ARCAF	173,4	8,8	9,6
Qiao Xing Universal Tele.	OTC	XING	64,3	22,1	30,8
Energy West Incorporated	OTC	EWST	29,1	9,7	16,3
Barnwell Industries, Inc.	AMEX	BRN	27,3	7,4	73,4
Innodata Corporation	OTC	INOD	66,1	11,0	29,6
Medical Action Industries	OTC	MDCI	137,1	26,9	30,6
Instrumentarium Corp.	OTC	INMRY	240,9	3,6	52,1
Petroleum Development	OTC	PETD	95,9	6,1	19,4
Drexler Technology Corp.	OTC	DRXR	233,6	45,6	53,6
Gerber Childrenswear Inc.	NYSE	GCW	126,9	7,9	25,8
Gaiam, Inc.	OTC	GAIA	295,5	68,2	60,7
Artesian Resources Corp.	OTC	ARTNA	62,8	20,5	45,5
York Water Company	OTC	YORW	92,2	22,9	74,2

pas suivies de façon détaillée par les analystes de Wall Street. Les données sont disponibles en ligne dans le fichier Marché secondaire.

a) Combien de variables y a-t-il dans l'ensemble de données ?

b) Lesquelles sont qualitatives ? Lesquelles sont quantitatives ?

c) Pour la variable Place boursière, calculer la fréquence et la fréquence en pourcentage pour le marché AMEX, la bourse de New York et le marché OTC. Construire un graphique en barres similaire à celui présenté à la figure 1.5 pour la variable Place boursière.

d) Déterminer la distribution de fréquence pour la marge brute en utilisant cinq intervalles : 0-14,9 ; 15-29,9 ; 30-44,9 ; 45-59,9 ; 60-74,9. Construire un histogramme similaire à la figure 1.6.

e) Quel est le coefficient de capitalisation boursière moyen ?

ANNEXE 1.1 UNE INTRODUCTION À STATTOOLS

StatTools est un module professionnel qui étend les capacités statistiques de Microsoft Excel.

Excel ne contient pas toutes les fonctions statistiques ou tous les outils d'analyse des données qui permettent d'effectuer l'ensemble des procédures statistiques décrites dans cet ouvrage. StatTools est un complément statistique à Microsoft Excel qui étend l'éventail des possibilités statistiques et graphiques d'Excel. La plupart des chapitres comprennent une annexe qui indique la démarche à suivre pour utiliser StatTools. Pour les étudiants qui souhaitent utiliser de façon plus approfondie le logiciel, StatTools offre un excellent système d'aide. Ce système d'aide inclut des explications détaillées des options d'analyse statistique et des données disponibles, ainsi que des descriptions et des définitions des types de résultats fournis.

A1.1.1 Débuter avec StatTools

Après avoir installé le logiciel, effectuez les étapes suivantes pour utiliser StatTools comme un module d'Excel.

Étape 1. Cliquez sur le bouton **Start** de la barre des tâches et cliquez sur **All Programs**.
Étape 2. Cliquez sur le fichier intitulé **Palisade Decision Tools**.
Étape 3. Cliquez sur **StatTools for Excel**.

Ces étapes entraîneront l'ouverture d'Excel et ajouteront StatTools dans le bandeau Excel. Si vous travaillez déjà avec Excel, ces étapes rendront StatTools disponible.

A1.1.2 Utiliser StatTools

Avant de commencer toute analyse statistique, vous devez créer un ensemble de données StatTools en utilisant le gestionnaire d'ensembles de données de StatTools. Utilisez la feuille Excel sur laquelle apparaissent les données sur les 60 pays de l'Organisation mondiale du comerce (tableau 1.1) pour illustrer ce que ça donne. Les étapes suivantes montrent comment créer un ensemble de données StatTools pour les données sur les 60 pays de l'OMC.

Étape 1. Ouvrir le fichier Excel appelé Nations.
Étape 2. Sélectionner une cellule dans l'ensemble de données (par exemple, la cellule A1).
Étape 3. Cliquez sur le bouton **StatTools** dans la barre des tâches.
Étape 4. Dans le groupe **Data**, cliquez sur **Data Set Manager**.
Étape 5. Lorsque StatTools demande si vous voulez ajouter le champ $A$1:$F$61 à un nouvel ensemble de données StatTools, cliquez sur **Yes**.
Étape 6. Lorsque la boîte de dialogue StatTools-Data Set Manager apparaît, cliquez sur **OK**.

La figure 1.10 montre la boîte de dialogue StatTools-Data Set Manager qui apparaît à l'étape 6. Par défaut, le nom du nouvel ensemble de données StatTools est Data Set #1. Vous pouvez remplacer le nom Data Set #1 dans l'étape 6 par un nom plus approprié.

A1.1.3 Applications recommandées

StatTools permet à l'utilisateur de spécifier l'endroit où les résultats seront affichés, ou comment les calculs seront effectués. Les étapes suivantes montrent comment accéder à la boîte de dialogue StatTools-Application Settings.

Étape 1. Cliquez sur le bouton **StatTools** dans la barre des tâches
Étape 2. Dans **Tools Group**, cliquez sur **Utilities**
Étape 3. Choisissez **Application Settings** dans la liste d'options

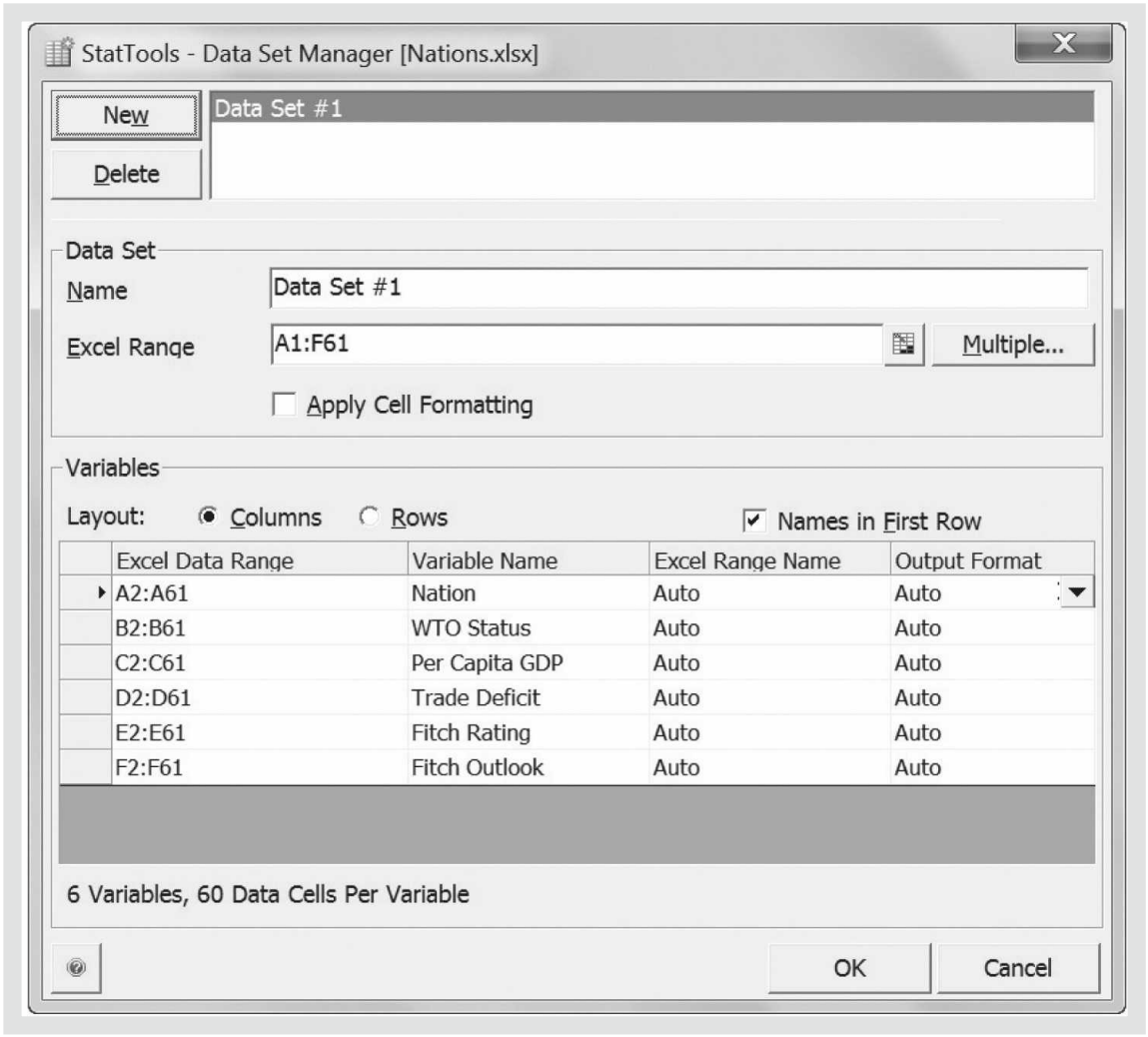


Figure 1.10 *La boîte de dialogue StatTools-Data Set Manager*

La figure 1.11 montre les cinq éléments de la boîte de dialogue StatTools-Application Settings : General Settings ; Reports ; Utilities ; Data Set Defaults et Analyses. Ci-dessous, nous montrons comment faire des changements dans la partie Reports de la boîte de dialogue.

La figure 1.11 indique que l'option Placement actuellement sélectionnée est **New Workbook**. En utilisant cette option, le résultat de StatTools sera placé dans un nouveau fichier. Mais supposez que vous vouliez placer le résultat dans le fichier actuellement actif. Si vous cliquez sur les mots **New Workbook**, une flèche pointée vers le bas apparaîtra à droite. En cliquant sur cette flèche, une liste de tous les emplacements possibles apparaîtra, dont **Active Workbook** ; nous recommandons d'utiliser cette option. La figure 1.11 révèle aussi que l'option **Updating Preferences** dans la partie Reports est actuellement **Live-Linked to Input Data**. Avec une mise à jour permanente, à chaque fois qu'une valeur est modifiée, StatTools changera automatiquement le résultat précédemment produit ; nous

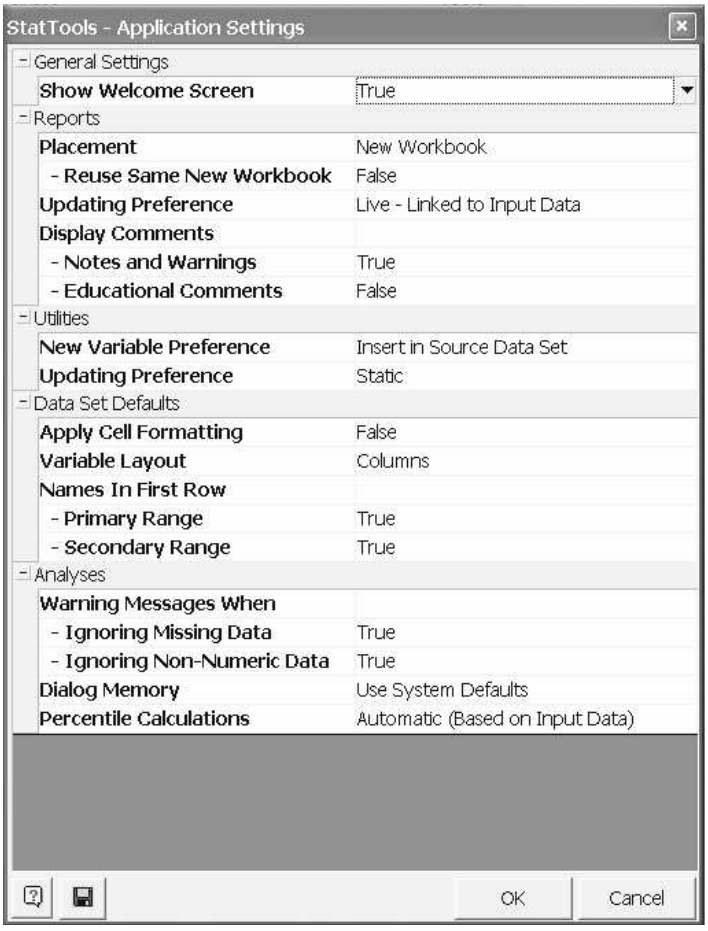


Figure 1.11 *La boîte de dialogue StatTools-Application Settings*

recommandons également d'utiliser cette option. Notez qu'il y a deux options disponibles sous **Display Comments** : **Notes and Warnings** et **Educational Comments**. Puisque ces options fournissent des informations utiles concernant le résultat, nous recommandons d'utiliser ces deux options. Ainsi, pour inclure des commentaires instructifs dans l'output de StatTools, vous devez modifier la valeur **False** par **True**.

La boîte de dialogue StatTools-Application Settings contient de nombreuses autres options qui vous permettent de personnaliser la façon dont vous souhaitez que StatTools opère. Vous pouvez en apprendre plus en sélectionnant l'option Aide située dans les outils ou en cliquant sur l'icône Aide de la boîte de dialogue. Lorsque vous avez fini de modifier les applications, cliquez sur OK en bas de la boîte de dialogue et ensuite cliquez sur Yes lorsque StatTools vous demande si vous souhaitez sauvegarder ces changements.