2

STATISTIQUES DESCRIPTIVES : PRÉSENTATIONS SOUS FORME DE TABLEAUX ET DE GRAPHIQUES

STATISTIQUES APPLIQUÉES

*La société Colgate-Palmolive**
New York, État de New York

La société Colgate-Palmolive est née d'un petit magasin de savons et de bougies, construit à New York en 1806. Aujourd'hui, Colgate-Palmolive emploie plus de 40 000 personnes dans plus de 200 pays à travers le monde. Bien que très connue pour ses produits de marque Colgate, Palmolive, Ajax et Fab, la société vend également les produits Mennen et les produits diététiques Hill.

La société Colgate-Palmolive utilise les instruments statistiques pour contrôler la qualité de ses produits lessive. Un des objectifs de ces programmes est de satisfaire les clients en contrôlant la quantité de lessive contenue dans un baril. Dans une catégorie de taille donnée, tous les barils sont remplis avec le même poids de poudre. Toutefois, le volume de poudre varie selon la densité de celle-ci. Par exemple, si la poudre est dense, un plus petit volume de détergent sera nécessaire pour obtenir le poids désiré. Par conséquent, un consommateur peut penser, en ouvrant le baril, que celui-ci n'est pas assez rempli.

Pour résoudre ce problème des poudres à forte densité, des densités limites ont été instaurées. Périodiquement, des échantillons de barils de lessive sont sélectionnés aléatoirement et la densité de la poudre de chaque échantillon est mesurée. Au vu des résultats, les responsables de la fabrication prennent les mesures qui s'imposent, afin de maintenir la densité dans les limites fixées.

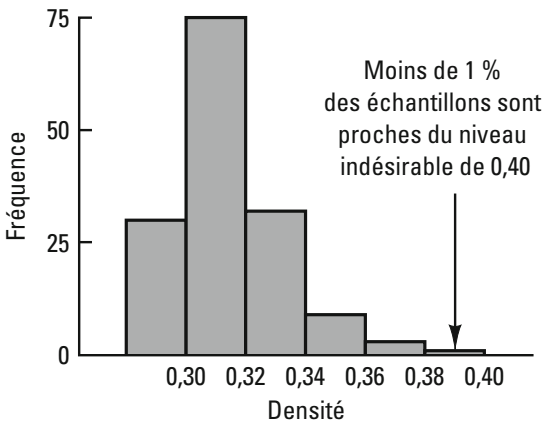
Une distribution de fréquence de la densité de 150 échantillons sélectionnés au cours d'une semaine et l'histogramme correspondant sont présentés ci-contre. Les densités supérieures à 0,4 sont jugées trop élevées. La distribution de fréquence et l'histogramme indiquent que les directives en matière de qualité sont respectées, toutes les densités étant inférieures ou égales à 0,4. Les managers, au regard de ces statistiques, peuvent être satisfaits de la qualité du processus de production.

Dans ce chapitre, nous étudierons les méthodes graphiques et les tableaux de statistiques descriptives, telles que les distributions de fréquence, les diagrammes en barres, les histogrammes, les diagrammes « stem-and-leaf », les tabulations croisées, etc. L'objectif de ces méthodes est de résumer les données de façon à pouvoir les comprendre et les interpréter plus facilement.

Distribution de fréquence des données sur la densité

Densité	Fréquence
0,29-0,30	30
0,31-0,32	75
0,33-0,34	32
0,35-0,36	9
0,37-0,38	3
0,39-0,40	1
Total	150

Histogramme des données sur la densité



* Les auteurs remercient William R. Fowle, responsable du département contrôle de la qualité chez Colgate-Palmolive, de leur avoir fourni ce Statistiques appliquées.

Comme nous l'avons vu au chapitre 1, les données peuvent être qualitatives (catégorielles) ou quantitatives. Les données qualitatives utilisent des labels ou des noms pour identifier différentes catégories d'une même variable. Les données quantitatives sont des valeurs numériques indiquant la quantité ou le nombre d'observations. Ce chapitre introduit les procédures graphiques et sous forme de tableaux habituellement utilisées pour décrire et résumer à la fois des données qualitatives et quantitatives. On trouve de telles descriptions dans des rapports annuels, des articles de journaux et des études. Tout le monde y est confronté. Par conséquent, il est important de comprendre comment elles sont élaborées et de savoir les interpréter correctement.

Nous commençons par les méthodes graphiques et sous forme de tableaux utilisées pour décrire des données concernant une seule variable. Nous introduisons ensuite les méthodes utilisées pour décrire des données relatives à deux variables et qui permettent d'établir la relation qui existe entre ces deux variables. La visualisation des données est un terme souvent utilisé pour décrire l'usage de graphiques pour résumer et présenter l'information contenue dans un ensemble de données. La dernière section de ce chapitre est une introduction à la visualisation des données et fournit quelques conseils pour créer des graphiques pertinents.

Les logiciels statistiques modernes étendent les capacités de description et de représentation graphique des données. Minitab et Excel sont deux logiciels assez répandus. Dans les annexes de ce chapitre, nous détaillerons certaines des possibilités offertes par ces logiciels.

2.1 RÉSUMER DES DONNÉES QUALITATIVES

2.1.1 Distribution de fréquence

Nous commençons notre discussion à propos de l'utilisation de graphiques et de tableaux dans le but de résumer des données qualitatives, en définissant une distribution de fréquence.

► **Distribution de fréquence**
Une distribution de fréquence est un résumé des données sous forme de tableau décrivant le nombre (la fréquence) des observations dans différentes classes juxtaposées.

Pour illustrer la construction et l'interprétation d'une distribution de fréquence pour des données qualitatives, considérons l'exemple suivant. Coca-Cola, Coca Light, Dr Pepper, Pepsi et Sprite sont cinq boissons non-alcoolisées largement répandues, consommées à travers le monde. Supposons que les données présentées dans le tableau 2.1 constituent un échantillon de 50 achats de boisson non-alcoolisée (fichier en ligne Boissons non alcoolisées).

Tableau 2.1 *Données issues d'un échantillon de 50 achats de boisson non-alcoolisée*

Coca-Cola	Coca Light	Pepsi
Coca Light	Coca-Cola	Dr. Pepper
Pepsi	Coca Light	Coca Light
Coca Light	Coca-Cola	Coca Light
Coca-Cola	Sprite	Pepsi
Coca-Cola	Pepsi	Pepsi
Dr. Pepper	Coca-Cola	Pepsi
Coca Light	Coca-Cola	Pepsi
Pepsi	Coca-Cola	Coca-Cola
Pepsi	Pepsi	Dr. Pepper
Coca-Cola	Coca-Cola	Pepsi
Dr. Pepper	Sprite	Sprite
Sprite	Dr. Pepper	
Coca-Cola	Pepsi	
Coca Light	Coca Light	
Coca-Cola	Pepsi	
Coca-Cola	Coca-Cola	
Sprite	Coca-Cola	
Coca-Cola	Coca-Cola	

Boissons non alcoolisées

Pour développer une distribution de fréquence à partir de ces données, le nombre de fois où chaque marque de boisson apparaît dans le tableau 2.1, est comptabilisé. Coca-Cola apparaît 19 fois, Coca Light 8 fois, Dr Pepper 5 fois, Pepsi 13 fois et Sprite 5 fois. Ces chiffres forment la distribution de fréquence présentée dans le tableau 2.2.

Cette distribution de fréquence résume la répartition des 50 achats de boisson entre les cinq marques. Ce résumé offre un aperçu plus pertinent des données que l'ensemble de données original, reproduit dans le tableau 2.1. D'après cette distribution de fréquence,

Tableau 2.2 *Distribution de fréquence des achats de boisson non-alcoolisée*

Boisson non-alcoolisée	Fréquence
Coca-Cola	19
Coca Light	8
Dr Pepper	5
Pepsi	13
Sprite	5
Total	50

Coca-Cola est le leader des ventes de boisson non-alcoolisée, Pepsi arrive en deuxième position, Coca Light en troisième position, Sprite et Dr Pepper occupent la quatrième place à égalité. La distribution de fréquence résume les informations sur la popularité des cinq marques de boisson non-alcoolisée les plus vendues.

2.1.2 Distributions de fréquence relative et en pourcentage

Une distribution de fréquence indique le nombre (la fréquence) d'observations dans chaque classe. Cependant, on s'intéresse souvent à la proportion ou au pourcentage d'observations dans chaque classe. La *fréquence relative* d'une classe correspond à la proportion des observations appartenant à cette classe. Pour un ensemble de données constitué de n observations, la fréquence relative de chaque classe est donnée par la relation suivante :

► **Fréquence relative**

$$\text{Fréquence relative d'une classe} = \frac{\text{Fréquence d'une classe}}{n} \qquad (2.1)$$

La *fréquence en pourcentage* d'une classe correspond à la fréquence relative multipliée par 100.

Une **distribution de fréquence relative** résume les données sous forme de tableau, en décrivant la fréquence relative de chaque classe. Une **distribution de fréquence en pourcentage** décrit la fréquence en pourcentage des données appartenant à chacune des classes. Le tableau 2.3 présente les distributions de fréquence relative et en pourcentage des données relatives aux achats de boisson non-alcoolisée. Dans le tableau 2.3, nous voyons que la fréquence relative pour Coca-Cola est de 19/50, soit 0,38 ; la fréquence relative pour Coca Light est égale à 8/50, soit 0,16 ; etc. Sur la base de la distribution de fréquence en pourcentage, on constate que 38 % des achats portent sur la marque Coca-Cola, 16 % sur la marque Coca Light, etc. On peut également remarquer que les trois premières marques représentent 80 % (38+26+16) des parts de marché.

Tableau 2.3 ***Distributions de fréquence relative et en pourcentage des achats de boisson non-alcoolisée***

Boisson non-alcoolisée	Fréquence relative	Fréquence en pourcentage
Coca-Cola	0,38	38
Coca Light	0,16	16
Dr Pepper	0,10	10
Pepsi	0,26	26
Sprite	0,10	10
Total	1,00	100

2.1.3 Diagramme en barres et diagramme circulaire

Un **diagramme en barres** est un moyen graphique de décrire des données qualitatives résumées par une distribution de fréquence absolue, relative ou en pourcentage. Sur l'un des axes du graphique (généralement l'axe horizontal), on note les labels ou noms utilisés pour identifier les classes (les catégories). Sur l'autre axe du graphique (généralement l'axe vertical), on note la fréquence absolue, relative ou en pourcentage. Chaque classe est représentée par une barre de largeur égale dont la hauteur correspond à la fréquence absolue, relative ou en pourcentage de la classe. Pour des données qualitatives, les barres doivent être séparées, reflétant le fait que chaque classe est une catégorie à part. La figure 2.1 représente le diagramme en barres de la distribution de fréquence des 50 achats de boisson non-alcoolisée. Le graphique révèle également que Coca-Cola, Pepsi et Coca Light sont les marques les plus achetées.

Dans les applications de contrôle de la qualité, les diagrammes en barres sont utilisés pour identifier les principales causes d'un problème. Lorsque les barres sont disposées en ordre décroissant, de gauche à droite, en fonction de leur hauteur, la cause la plus fréquente apparaît alors en premier. Ce type de diagramme en barres est appelé diagramme de Pareto, du nom de son inventeur, Vilfredo Pareto, un économiste italien.

Le **diagramme circulaire** est un autre type de graphique permettant de représenter les distributions de fréquence relative et en pourcentage de données qualitatives. Pour dessiner un diagramme circulaire, il faut tout d'abord tracer un cercle représentant l'ensemble des données. Ensuite, on se sert des fréquences relatives pour diviser le cercle en secteurs, ou parts, qui correspondent à la fréquence relative de chaque classe. Par exemple, puisqu'un cercle fait 360 degrés et que la marque Coca-Cola a

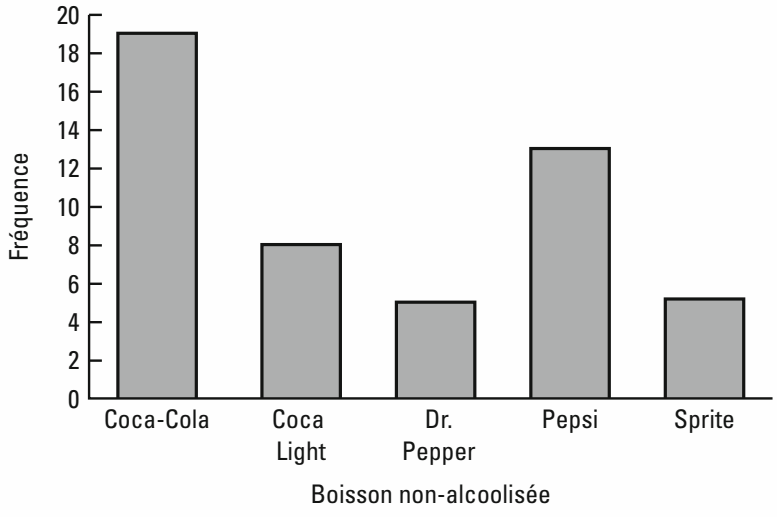


Figure 2.1 *Diagramme en barres des achats de boisson non-alcoolisée*

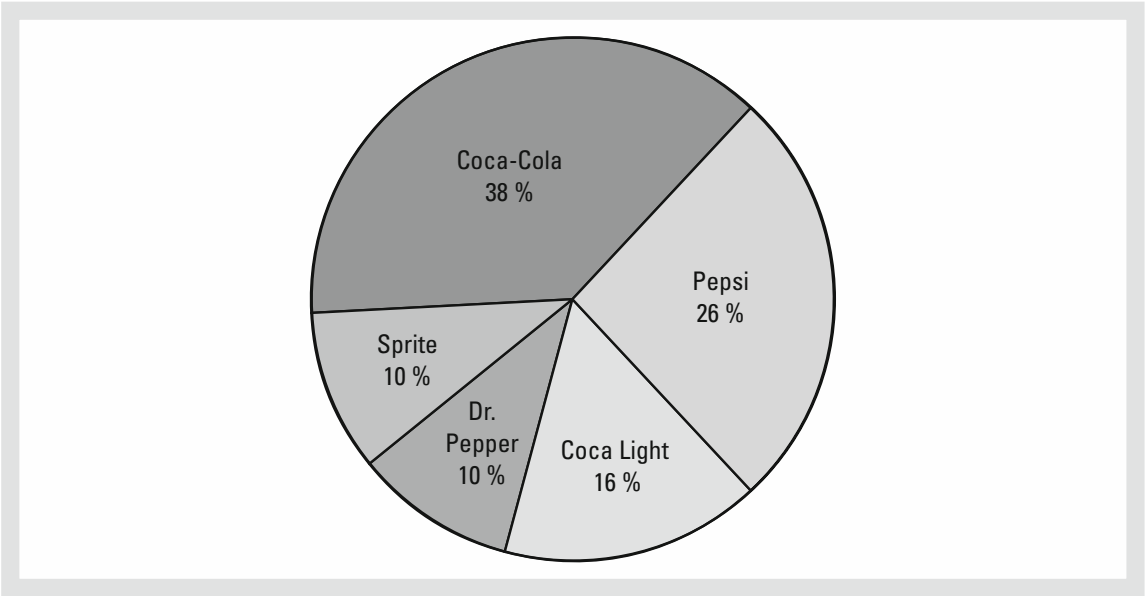


Figure 2.2 ***Diagramme circulaire des achats de boisson non-alcoolisée***

une fréquence relative de 0,38, le secteur du diagramme circulaire correspondant à la marque Coca-Cola fait 136,8 degrés (0,38 × 360 = 136,8). Le secteur du diagramme circulaire correspondant à la marque Coca Light fait 57,6 degrés (0,16 × 360 = 57,6). Des calculs similaires pour les autres classes permettent de construire le diagramme circulaire de la figure 2.2. Les valeurs numériques utilisées pour déterminer l'angle de chaque secteur peuvent être indifféremment les fréquences absolues, relatives ou en pourcentage.

De multiples options dans le choix des couleurs et des hachures, dans la disposition de la légende, du titre et la possibilité de représenter le graphique en trois dimensions, améliorent l'apparence visuelle des diagrammes en barres et circulaires. Lorsqu'elles sont correctement utilisées, ces options permettent d'obtenir un graphique plus pertinent. Mais ce n'est pas toujours le cas. Considérez par exemple le diagramme circulaire pour les boissons non-alcoolisées en trois dimensions représenté à la figure 2.3. Comparez-le à la représentation plus simple présentée à la figure 2.2. La perspective en trois dimensions n'apporte rien à la compréhension du graphique. En réalité, dans la mesure où la perspective en trois dimensions nous oblige à visualiser le diagramme circulaire de la figure 2.3 sous un certain angle plutôt qu'à plat, la visualisation des données est plus complexe. L'utilisation d'une légende dans la figure 2.3 vous oblige à reporter sans cesse votre regard de la légende au diagramme. Le graphique plus simple représenté à la figure 2.2, qui indique les pourcentages et les catégories directement sur le diagramme circulaire, est plus efficace.

En général, les diagrammes circulaires ne sont pas la meilleure façon de représenter des pourcentages à comparer. Les recherches ont prouvé que les individus appréhendent plus facilement des différences représentées par des longueurs différentes

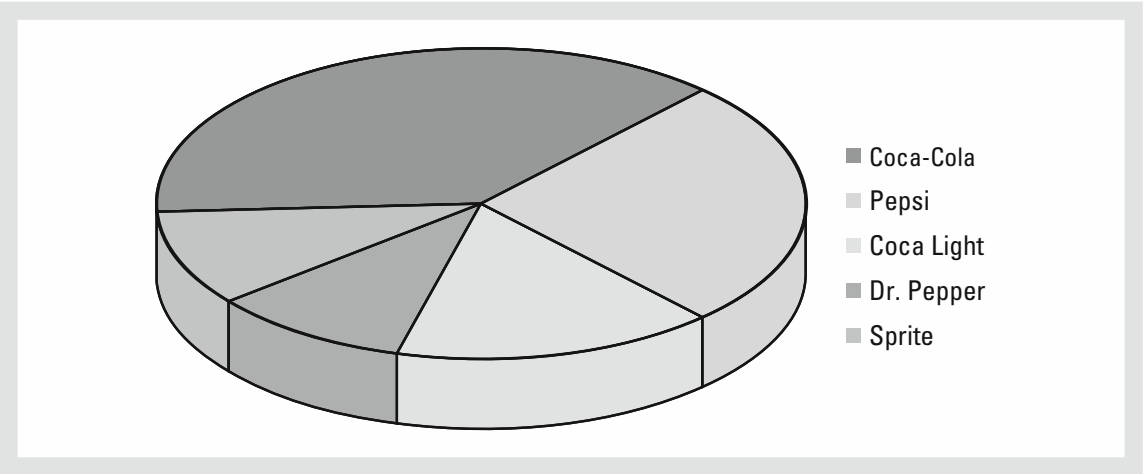


Figure 2.3 ***Diagramme circulaire en trois dimensions pour les achats de boisson non-alcoolisée***

que par des sections (ou des parts) différentes. Pour faire de telles comparaisons, nous recommandons l'utilisation de diagrammes en barres similaires à celui de la figure 2.1. Dans la section 2.5, nous fournirons de plus amples conseils pour créer des graphiques pertinents.

REMARQUES

1. Souvent, le nombre de classes d'une distribution de fréquence correspond au nombre de catégories définies parmi les données, comme c'est le cas pour les données concernant les achats de boisson non-alcoolisée dans cette section. Les données concernent cinq marques de boisson et la distribution de fréquence comprend cinq classes, représentant ces cinq marques. Des données qui incluraient toutes les marques de boisson non-alcoolisée existantes sur le marché, comporteraient de nombreuses catégories, beaucoup n'ayant qu'un nombre total d'achats très faible. La plupart des statisticiens recommandent de regrouper ces classes, caractérisées par de faibles fréquences, en une seule classe agrégée, désignée par le terme « autre ». Les classes dont les fréquences sont inférieures ou égales à 5 %, seront généralement regroupées.
2. La somme des fréquences dans une distribution de fréquence est toujours égale au nombre d'observations. La somme des fréquences relatives dans une distribution de fréquence relative est toujours égale à 1 et la somme des pourcentages dans une distribution de fréquence en pourcentage est toujours égale à 100.

EXERCICES

Méthode

1. Trois réponses à une question sont possibles : A, B et C. Un échantillon de 120 réponses fournit 60 A, 24 B et 36 C. Donner les distributions de fréquence absolue et relative.

2. Une partie d'une distribution de fréquence relative est donnée ci-dessous.

Classe	Fréquence relative
A	0,22
B	0,18
C	0,40
D	

a) Quelle est la fréquence relative de la classe D ?

b) La taille de l'échantillon est égale à 200. Quelle est la fréquence de la classe D ?

c) Donner la distribution de fréquence.

d) Donner la distribution de fréquence en pourcentage.

3. Les réponses à un questionnaire sont les suivantes : 58 oui, 42 non et 20 sans opinion.

a) Dans un diagramme circulaire, combien de degrés aurait la section représentant les réponses positives ?

b) Combien de degrés aurait la section du diagramme représentant les réponses négatives ?

c) Construire un diagramme circulaire.

d) Construire un diagramme en barres.

Applications

4. Lors de la saison 2010-2011, les cinq programmes télévisés les plus regardés étaient *la Roue de la Fortune* (RF), *Deux hommes et demi* (DHD), *Jeopardy* (Jep), le *Juge Judy* (JJ) et le *Show d'Oprah Winfrey* (SOW) (site Internet de Nielsen Media Research, 16 avril 2012). Les données indiquant les émissions préférées d'un échantillon de 50 téléspectateurs sont fournies ci-dessous (fichier en ligne Émissions).

RF	DHD	Jep
DHD	DHD	JJ
Jep	DHD	RF
RF	JJ	JJ
DHD	SOW	Jep
SOW	RF	SOW
JJ	SOW	DHD

DHD	JJ	DHD
Jep	JJ	RF
RF	DHD	RF
Jep	RF	DHD
RF	RF	Jep
SOW	SOW	RF
DHD	Jep	JJ
JJ	Jep	Jep
SOW	RF	Jep
RF	DHD	

a) Ces données sont-elles qualitatives ou quantitatives ?

b) Donner les distributions de fréquence absolue et en pourcentage de ces données.

c) Construire un diagramme en barres et un diagramme circulaire.

d) En se basant sur les données de l'échantillon, quelle émission a eu la plus grande audience ? Quelle est la seconde ?

5. Par ordre alphabétique, les six noms de famille les plus courants aux États-Unis sont Brown, Johnson, Jones, Miller, Smith et Williams (*The World Almanac*, 2012). Supposez qu'un échantillon de 50 individus dont le nom de famille correspond à l'un de ces six noms, fournisse les données suivantes (fichier en ligne Nom de famille 2012) :

Brown	Williams	Williams	Williams	Brown
Smith	Jones	Smith	Johnson	Smith
Miller	Smith	Brown	Williams	Johnson
Johnson	Smith	Smith	Johnson	Brown
Williams	Miller	Johnson	Williams	Johnson
Williams	Johnson	Jones	Smith	Brown
Johnson	Smith	Smith	Brown	Jones
Jones	Jones	Smith	Smith	Miller
Miller	Jones	Williams	Miller	Smith
Jones	Johnson	Brown	Johnson	Miller

Résumer les données en construisant :

a) Les distributions de fréquence relative et en pourcentage

b) Un digramme en barres

c) Un diagramme circulaire

d) En vous basant sur ces données, quels sont les trois noms de famille les plus courants ?

6. L'institut Nielsen Media Research a fourni la liste des 25 programmes les mieux notés de l'histoire de la télévision (*The World Almanac*, 2012). Les données suivantes indiquent la chaîne de télévision qui a produit chacun de ces 25 programmes (fichier en ligne Chaîne).

CBS	CBS	NBC	FOX	CBS
CBS	NBC	NBC	NBC	ABC
ABC	NBC	ABC	ABC	NBC
CBS	NBC	CBS	ABC	NBC
NBC	CBS	CBS	ABC	CBS

a) Construire une distribution de fréquence, de fréquence en pourcentage et un diagramme en barres pour ces données.

b) Quelle(s) chaîne(s) a (ont) présenté le plus de programmes les mieux notés ? Comparer les performances des chaînes ABC, CBS et NBC.

7. L'enquête de satisfaction des clients des aéroports menée par le centre de recherche Canmark utilise un questionnaire en ligne pour donner aux compagnies aériennes et aux aéroports des informations sur les taux de satisfaction des clients, relatifs à divers éléments de leur vol (site Internet Airport Survey, juillet 2012). Après avoir effectué un vol, les clients reçoivent un e-mail leur demandant d'aller sur le site Internet et de noter divers facteurs dont le processus de réservation, le processus d'enregistrement, la politique concernant les bagages, la propreté de l'aire d'embarquement, le service offert par les hôtesses, la variété des plats et des boissons proposés, la ponctualité, etc. Une échelle de notation comprenant 5 niveaux (Excellent (E), Très bon (T), Bon (B), Convenable (C) et Mauvais (M)) est utilisée pour enregistrer les notes octroyées par les clients à chaque item. Supposez que les passagers d'un vol Delta Airlines en partance de Myrtle Beach, en Caroline du Sud et à destination d'Atlanta en Géorgie, aient fourni les évaluations suivantes à la question : « S'il vous plaît, noter la compagnie en fonction de votre expérience globale lors de ce vol ». Les évaluations sont les suivantes (fichier en ligne Enquête aérienne) :

E	E	B	T	T	E	T	T	T	E
E	B	T	E	E	T	E	E	E	T
T	T	T	C	T	E	T	E	B	E
B	E	T	E	T	E	T	T	T	T
E	E	T	T	E	M	E	T	M	T

a) Utilisez une distribution de fréquence en pourcentage et un diagramme en barres pour résumer ces données. Qu'indiquent ces résumés à propos de la satisfaction globale des clients de ce vol Delta Airlines ?

b) Le questionnaire en ligne permet aux personnes interrogées de s'exprimer librement à propos des éventuels problèmes rencontrés. Est-ce que cela est une information utile pour un responsable qui cherche à améliorer la satisfaction globale des clients des vols Delta Airline ? Expliquez.

8. Les positions d'un échantillon de 55 membres du club de baseball Hall of Fame de Cooperstown, dans l'État de New York, sont présentées ci-dessous (fichier en ligne Baseball Hall). Chaque observation indique la position principale occupée par les Hall of Famers : lanceur (L), receveur (R), 1ère base (1), 2e base (2), 3e base (3), bloqueur (B), champ gauche (G), champ droit (D) et milieu de terrain (M).

G	R	M	L	2	R	1	B	B	1	G	R
R	R	R	D	M	G	D	R	M	M	R	R
2	3	R	L	G	1	M	R	R	R	B	1
D	1	2	L	B	L	2	G	R	D	D	G
R	R	D									

a) Utiliser les distributions de fréquence absolue et relative pour résumer les données.

b) Quelle est la position la plus occupée par les Hall of Famers ?

c) Quelle est la position la moins occupée par les Hall of Famers ?

d) Quelle est la position hors jeu (G, M ou D) la plus occupée par les Hall of Farmers ?

e) Comparer les joueurs dans le champ (1, 2, 3 et B) et les joueurs hors champ (G, M, D).

9. L'étude du centre de recherche Pew sur les tendances démographiques et sociales a conclu que 46 % des adultes américains aimeraient vivre dans un endroit différent de celui dans lequel ils vivent actuellement (Centre de recherche Pew, 29 janvier 2009). L'enquête nationale réalisée auprès de 2 260 adultes posait les questions suivantes « Où vivez-vous ? » et « Quel est l'endroit idéal selon vous ? ». Les réponses possibles étaient Ville (V), Banlieue (B), Petite ville (P) et Zone rurale (R). Les réponses fournies par un échantillon représentatif de 100 personnes sont présentées ci-dessous (fichier en ligne Zone d'habitation).

Zone d'habitation

Où vivez-vous aujourd'hui ?

B	P	R	V	R	R	P	V	B	P
V	B	V	B	P	B	B	V	B	B
P	P	V	V	B	P	V	B	P	V
P	R	B	B	P	V	B	V	P	V
P	V	P	V	R	V	V	R	P	V
B	B	P	B	V	V	V	R	B	V
B	B	V	V	B	V	R	P	P	P
V	R	P	V	R	V	P	R	R	V
P	V	V	R	P	P	R	B	R	P
P	B	B	B	B	B	V	V	R	P

Quel est l'endroit idéal selon vous ?

B	V	R	R	R	B	P	B	B	P
P	B	V	B	P	V	V	R	P	R
C	P	P	B	B	V	V	P	P	B
B	R	V	B	V	V	B	V	R	V
P	B	R	R	R	V	P	B	P	P
P	R	R	B	V	V	R	R	B	B
B	P	V	P	P	V	R	P	P	P
V	P	P	R	R	V	B	R	P	V
P	V	V	P	P	P	R	V	R	P
P	V	B	B	V	B	P	B	B	R

a) Fournir une distribution de fréquence en pourcentage pour chaque question.

b) Construire un diagramme en barres pour chaque question.

c) Où vivent actuellement la plupart des adultes ?

d) Quel serait l'endroit idéal pour la plupart des adultes ?

e) Quels changements dans les zones d'habitation vous attendriez-vous à voir si les gens quittaient leur lieu d'habitation actuel pour aller vivre dans leur lieu préféré ?

10. Virtual Tourist note les hôtels à travers le monde. Les notes fournies par 649 personnes ayant fréquenté l'hôtel Sheraton d'Anaheim, situé près de Disneyland Resort, en Californie, sont disponibles dans le fichier en ligne HotelRatings (site Internet de Virtual Tourist, 25 février 2013). Les réponses possibles étaient Excellent, Très bon, Convenable, Mauvais, Vraiment mauvais.

a) Construire une distribution de fréquence.

b) Construire une distribution de fréquence en pourcentage.

c) Construire un diagramme en barres pour la distribution de fréquence en pourcentage.

d) Comment les personnes ayant fréquenté l'hôtel Sheraton d'Anaheim évaluent-elles leur séjour ?

e) Les notes obtenues auprès de 1 679 personnes qui ont séjourné dans le Grand Californian de Disney sont résumées par la distribution de fréquence suivante :

Note	Fréquence
Excellente	807
Très bonne	521
Convenable	200
Mauvaise	107
Vraiment mauvaise	44

Comparez les notes obtenues par l'hôtel Grand Californian de Disney à celles obtenues par l'hôtel Sheraton d'Anaheim.

2.2 RÉSUMER DES DONNÉES QUANTITATIVES

2.2.1 Distribution de fréquence

Comme nous l'avons déjà dit dans la section 2.1, une distribution de fréquence est un résumé sous forme de tableau, décrivant le nombre (la fréquence) d'observations contenues dans chaque classe ou catégorie juxtaposée (qui ne se chevauchent pas). Cette définition reste valable pour des données quantitatives. Cependant, il convient d'être plus attentif à la définition des classes utilisées pour construire une distribution de fréquence lorsqu'il s'agit de données quantitatives.

Tableau 2.4 ***Durée (en jours) des audits de fin d'année***

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

Considérons par exemple les données quantitatives figurant dans le tableau 2.4. Ces données indiquent le temps nécessaire (en jours) pour effectuer l'audit de fin d'année de 20 clients de Sanderson et Clifford, un petit cabinet d'experts-comptables. Les trois étapes nécessaires à la définition des classes d'une distribution de fréquence pour des données quantitatives sont :

1. Déterminer le nombre de classes juxtaposées

2. Déterminer la largeur de la classe

3. Déterminer les limites de la classe

Illustrons ces étapes en développant une distribution de fréquence pour les données du tableau 2.4.

Nombre de classes – Les classes regroupent les observations en fonction de leurs caractéristiques. En général, on recommande d'utiliser entre 5 et 20 classes. Lorsque le nombre d'observations est relativement faible, cinq ou six classes suffisent généralement pour répartir les données. Pour un nombre plus important d'observations, un nombre plus important de classes est généralement nécessaire. L'objectif est d'utiliser suffisamment de classes pour souligner les divergences, ou différences qui existent entre les données, sans toutefois obtenir un nombre excessif de classes qui se traduirait par le fait que certaines classes ne seraient constituées que de quelques observations. Puisque l'ensemble de données du tableau 2.4 est relativement petit ($n = 20$), nous avons choisi de développer une distribution de fréquence en 5 classes.

Largeur des classes – La seconde étape dans la construction d'une distribution de fréquence pour des données quantitatives consiste à choisir la largeur des classes. Nous recommandons de choisir la même largeur pour toutes les classes. Ainsi, les choix du nombre de classes et de leur largeur ne sont pas indépendants. Plus le nombre de classes est important, moins la classe sera large et vice versa. Pour déterminer la largeur de classe appropriée, on identifie la plus petite et la plus grande valeur de l'ensemble de données. Ensuite, une fois le nombre de classes spécifié, on peut utiliser l'expression suivante pour déterminer la largeur approximative de la classe.

$$\text{Largeur approximative de la classe} = \frac{\text{Valeur la plus élevée} - \text{Valeur la plus faible}}{\text{Nombre de classes}} \qquad (2.2)$$

Utiliser la même largeur pour chaque classe réduit la probabilité que l'utilisateur interprète mal la distribution de fréquence.

La largeur approximative de la classe donnée par l'équation (2.2) peut être arrondie à une valeur plus appropriée, en fonction des préférences de la personne qui crée la distribution de fréquence. Par exemple, une largeur approximative de classe de 9,28 peut être arrondie à 10, simplement parce que 10 est une largeur de classe plus adéquate pour construire une distribution de fréquence.

Dans l'ensemble de données sur la durée des audits de fin d'année, la valeur la plus élevée est 33 et la plus petite est 12. Puisque nous avons décidé de répartir les données en 5 classes, la largeur approximative d'une classe est égale à 4,2 ($\left(33-12\right)\quad 5=4,2$), selon l'équation (2.2). Par conséquent, nous décidons d'arrondir ce chiffre et d'utiliser une largeur de classe de 5 jours pour construire la distribution de fréquence.

En pratique, le nombre de classes et la largeur approximative des classes sont déterminés par un processus d'essai-erreur. Lorsqu'un nombre de classes est choisi, l'équation (2.2) est utilisée pour trouver la largeur approximative de la classe. Le processus peut être répété pour un nombre de classes différent. Finalement, l'analyste fait appel à son bon sens pour déterminer la combinaison nombre de classes – largeur de classe qui fournit la distribution de fréquence la plus pertinente pour résumer les données.

> Aucune distribution de fréquence n'est meilleure qu'une autre pour un même ensemble de données. Des individus différents peuvent construire des distributions de fréquence différentes mais toutes acceptables. L'objectif est de révéler le regroupement naturel des données et les différences qui peuvent exister.

Après avoir décidé d'utiliser 5 classes, chacune d'une largeur de 5 jours pour construire la distribution de fréquence des données sur la durée des audits du tableau 2.4, l'étape suivante consiste à spécifier les limites de classe pour chacune de ces classes.

Limites de classe – Les limites de classe doivent être choisies de sorte à ce que chaque observation appartienne à une et une seule classe. La *limite inférieure de classe* identifie la plus petite valeur possible assignée à la classe. La *limite supérieure de classe* identifie la plus grande valeur possible assignée à la classe. Pour développer des distributions de fréquence pour des données qualitatives, nous n'avons pas besoin de spécifier les limites de classes car chaque observation appartient à une classe séparée. Mais avec des données quantitatives, comme la durée des audits du tableau 2.4, il est nécessaire de définir les limites de classe pour déterminer à quelle classe appartient chaque observation.

Pour les données sur la durée des audits du tableau 2.4, nous sélectionnons 10 jours comme étant la limite inférieure et 14 comme étant la limite supérieure de la première classe. Cette classe est notée 10-14 dans le tableau 2.5. La plus petite observation, 12, est incluse dans la classe 10-14. Nous sélectionnons ensuite 15 jours comme la limite inférieure et 19 la limite supérieure de la deuxième classe. Nous continuons ainsi et obtenons les cinq classes suivantes : 10-14, 15-19, 20-24, 25-29 et 30-34. La plus grande observation, 33, est incluse dans la classe 30-34. La différence entre les limites inférieures de deux classes adjacentes correspond à la largeur de la classe. En utilisant les deux premières limites inférieures de classe, 10 et 15, on constate que la largeur d'une classe est égale à 5 (15 – 10 = 5).

Tableau 2.5 *Distribution de fréquence pour les données sur la durée des audits*

Durée de l'audit (en jours)	Fréquence
10-14	4
15-19	8
20-24	5
25-29	2
30-34	1
Total	20

Une fois le nombre de classes fixé, leur largeur et leurs limites déterminées, une distribution de fréquence peut être obtenue en comptabilisant le nombre d'observations appartenant à chaque classe. Par exemple, quatre observations des données du tableau 2.4 (12, 14, 14 et 13) appartiennent à la classe 10-14. Ainsi, la fréquence de la classe 10-14 est 4. En poursuivant ce processus de comptabilisation pour les classes 15-19, 20-24, 25-29 et 30-34, on obtient la distribution de fréquence présentée dans le tableau 2.5. En utilisant cette distribution de fréquence, on observe que :

- Les durées d'audit les plus fréquemment observées appartiennent à la classe 15-19 jours. Huit audits sur vingt appartiennent à cette classe.
- Seul un audit a nécessité plus de 30 jours.

D'autres conclusions sont possibles, selon les centres d'intérêt de la personne qui examine la distribution de fréquence. L'intérêt d'une distribution de fréquence est de fournir des informations sur les données que l'on ne peut pas obtenir facilement à partir de l'ensemble de données original.

Centre d'une classe : Dans certaines applications, il est nécessaire de connaître le centre des classes d'une distribution de fréquence relative à des données quantitatives. Le **centre d'une classe** est la valeur médiane entre les limites inférieure et supérieure de classe. Pour les données sur la durée des audits, le centre des cinq classes est respectivement 12, 17, 22, 27 et 32.

2.2.2 Distributions de fréquence relative et en pourcentage

Nous définissons les distributions de fréquence relative et en pourcentage pour des données quantitatives de la même manière que pour des données qualitatives. Premièrement, rappelons que la fréquence relative est simplement la proportion des observations appartenant à une classe. Avec n observations,

$$\text{Fréquence relative d'une classe} = \frac{\text{Fréquence de cette classe}}{n}$$

La fréquence en pourcentage d'une classe est la fréquence relative multipliée par 100.

Tableau 2.6 ***Distributions de fréquence relative et en pourcentage pour les données sur la durée des audits***

Durée de l'audit (en jours)	Fréquence relative	Fréquence en pourcentage
10-14	0,20	20
15-19	0,40	40
20-24	0,25	25
25-29	0,10	10
30-34	0,05	5
Total	1,00	100

Basé sur la fréquence des classes du tableau 2.5, et avec $n = 20$, le tableau 2.6 présente les distributions de fréquence relative et en pourcentage des données relatives aux audits. Notez que 0,40, soit 40 % des audits nécessitent entre 15 et 19 jours. Seulement 0,05, soit 5 % des audits nécessitent au moins 30 jours. De nouveau, d'autres interprétations et informations peuvent être déduites du tableau 2.6.

2.2.3 Diagramme de points

L'un des résumés graphiques de données les plus simples est le diagramme de points. L'étendue des données est représentée sur un axe horizontal. Chaque observation est représentée par un point placé au-dessus de l'axe. La figure 2.4 correspond au diagramme de points des données sur la durée des audits du tableau 2.4. Les trois points placés au-dessus de la valeur 18 sur l'axe horizontal indiquent qu'à trois reprises, l'audit a duré 18 jours. Les diagrammes de points détaillent les données et sont utiles pour comparer la distribution de plusieurs variables.

2.2.4 Histogramme

Une autre représentation graphique courante des données quantitatives est l'histogramme. Ce graphique peut être réalisé à partir de données préalablement résumées par une distribution de fréquence absolue, relative ou en pourcentage. Un histogramme est construit en plaçant la variable considérée sur l'axe horizontal et la fréquence absolue,

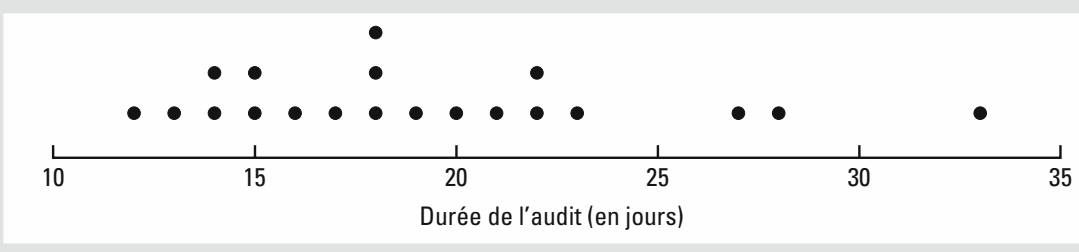


Figure 2.4 ***Diagramme de points pour les données sur la durée des audits***

relative ou en pourcentage sur l'axe vertical. La fréquence absolue, relative ou en pourcentage de chaque classe est représentée par un rectangle dont la base est déterminée par les limites de classes et dont la hauteur correspond à la fréquence absolue, relative ou en pourcentage.

La figure 2.5 représente un histogramme pour les données sur la durée des audits. Notez que la classe ayant la plus grande fréquence correspond à la classe 15-19 jours. La hauteur du rectangle au-dessus de cette classe révèle que la fréquence de cette classe est égale à 8. Un histogramme pour la distribution relative ou en pourcentage de ces données aurait la même forme, mis à part le fait que l'axe vertical représenterait les fréquences relatives ou en pourcentage.

Comme le montre la figure 2.5, les rectangles adjacents d'un histogramme se touchent. Contrairement à un diagramme en barres, un histogramme ne contient pas de séparation naturelle entre les rectangles des classes adjacentes. Cette présentation est la convention habituelle pour les histogrammes. Puisque les classes pour les données sur la durée des audits sont définies par les intervalles suivants 10-14, 15-19, 20-24, 25-29 et 30-34, un espace d'une unité (de 14 à 15, de 19 à 20, de 24 à 25, de 29 à 30) semble être nécessaire entre les classes. Ces espaces sont éliminés en construisant l'histogramme. L'élimination des espaces entre les classes d'un histogramme pour les données relatives à la durée des audits souligne le fait que toutes les valeurs comprises entre la limite inférieure de la première classe et la limite supérieure de la dernière classe sont possibles.

L'un des principaux attraits d'un histogramme est de fournir des informations concernant la forme d'une distribution. La figure 2.6 présente quatre histogrammes construits à partir de distributions de fréquence relative. Le cas A représente l'histogramme d'un ensemble de données modérément asymétrique ou biaisé à gauche. Un histogramme est dit asymétrique ou biaisé à gauche si sa queue de distribution s'étend vers la gauche. Ce type d'histogramme est caractéristique des résultats d'examens, aucune note n'étant

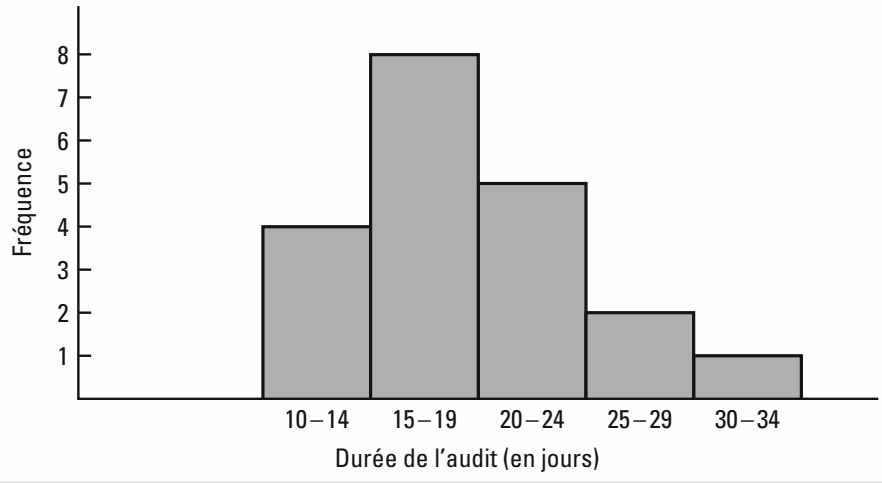


Figure 2.5 *Histogramme pour les données sur la durée des audits*

supérieure à 100 % de bonnes réponses, la plupart des notes étant supérieures à 70 %. Le cas B illustre l'histogramme d'un ensemble de données modérément asymétrique à droite. Un histogramme est dit asymétrique à droite si sa queue de distribution s'étend davantage à droite. Des données relatives aux prix des logements fournissent un exemple de ce type d'histogramme : quelques logements très chers créent une asymétrie dans la queue droite de la distribution.

Le cas C représente un histogramme symétrique. Dans un histogramme symétrique, les queues de distribution droite et gauche ont la même forme. Les histogrammes obtenus à partir de données réelles ne sont jamais parfaitement symétriques, mais peuvent l'être à peu près. Des données relatives à la taille ou au poids d'individus fournissent des histogrammes relativement symétriques. Le cas D illustre un histogramme fortement asymétrique à droite. Cet histogramme a été construit à partir de données relatives aux montants des achats des clientes d'un magasin d'habillement pour femme au cours d'une journée. Les données issues d'applications en économie conduisent souvent à des histogrammes asymétriques à droite. Par exemple, les données concernant les prix des logements, les salaires, les quantités achetées, etc. sont représentées par des histogrammes asymétriques à droite.

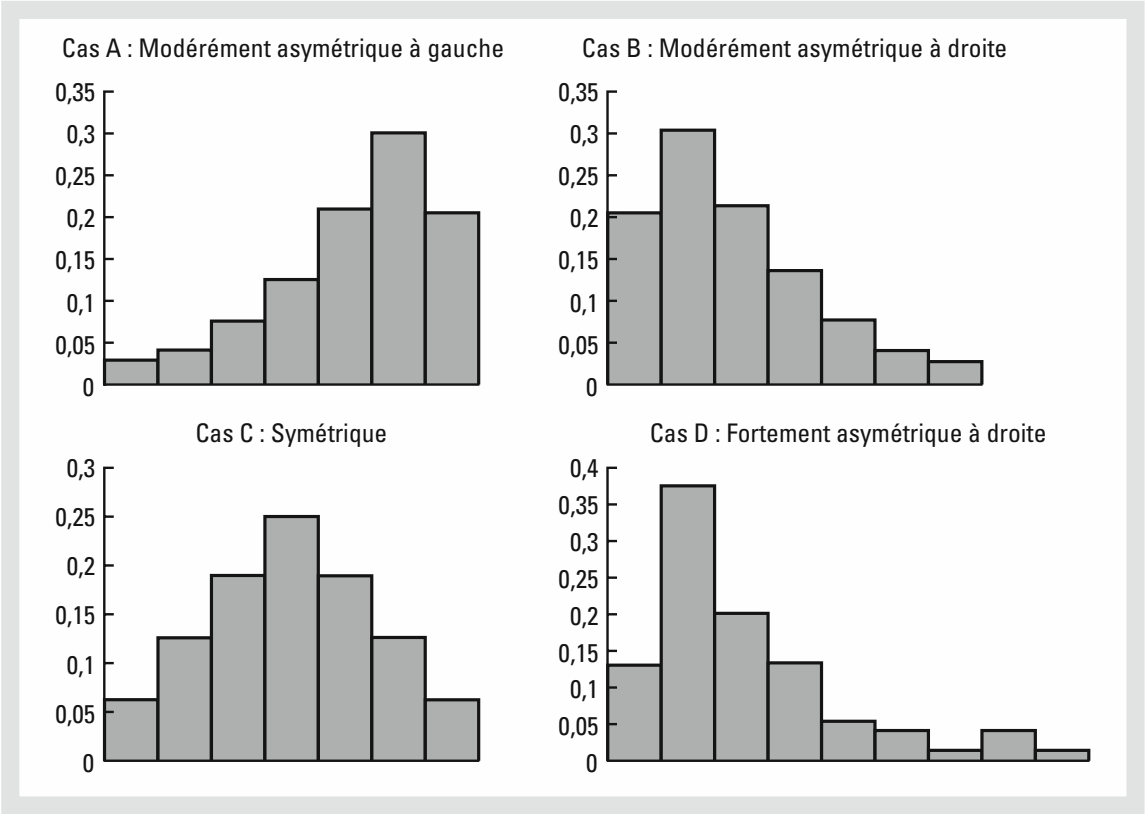


Figure 2.6 ***Histogrammes illustrant différents degrés d'asymétrie***

2.2.5 Distributions cumulées

Une variante de la distribution de fréquence qui fournit un autre résumé des données quantitatives, sous forme de tableau, est la **distribution de fréquence cumulée**. La distribution de fréquence cumulée utilise le nombre, la largeur et les limites des classes développées pour la distribution de fréquence. Cependant, plutôt que de représenter la fréquence de chaque classe, la distribution de fréquence cumulée représente le nombre d'observations dont les valeurs sont *inférieures ou égales à la limite supérieure de chaque classe*. Les deux premières colonnes du tableau 2.7 fournissent la distribution de fréquence cumulée des données sur la durée des audits.

Pour comprendre comment les fréquences cumulées sont calculées, considérons la classe intitulée « inférieure ou égale à 24 ». La fréquence cumulée de cette classe est simplement la somme des fréquences de toutes les classes dont les observations sont inférieures ou égales à 24. À partir de la distribution de fréquence du tableau 2.5, la somme des fréquences des classes 10-14, 15-19 et 20-24 indique qu'il y a 17 observations (4 + 8 + 5 = 17) dont la valeur est inférieure ou égale à 24. Par conséquent, la fréquence cumulée pour cette classe est égale à 17. De plus, la distribution de fréquence cumulée présentée dans le tableau 2.7 révèle que 4 audits ont été réalisés en 14 jours au maximum et 19 audits ont été réalisés en 29 jours au maximum.

Pour finir, notez qu'une **distribution de fréquence cumulée relative**, respectivement **en pourcentage**, fournit la proportion, respectivement le pourcentage, des observations dont la valeur est inférieure ou égale à la limite supérieure de chaque classe. La distribution de fréquence cumulée relative peut être calculée soit en sommant les fréquences relatives de la distribution de fréquence relative, soit en divisant les fréquences cumulées par le nombre total d'observations. Les fréquences cumulées relatives présentées dans la colonne 3 du tableau 2.7 ont été obtenues en divisant les fréquences cumulées de la colonne 2 par le nombre total d'observations ($n = 20$). Les fréquences cumulées en pourcentage ont été calculées en multipliant les fréquences cumulées relatives par 100. Les distributions de fréquence cumulée relative et en pourcentage montrent que 0,85, soit 85 % des audits ont été réalisés en moins de 25 jours, 0,95, soit 95 % des audits ont été réalisés en moins de 30 jours, etc.

Tableau 2.7 *Distributions de fréquence cumulée absolue, relative et en pourcentage pour les données sur la durée des audits*

Durée des audits (en jours)	Fréquence cumulée	Fréquence cumulée relative	Fréquence cumulée en pourcentage
Inférieure ou égale à 14	4	0,20	20
Inférieure ou égale à 19	12	0,60	60
Inférieure ou égale à 24	17	0,85	85
Inférieure ou égale à 29	19	0,95	95
Inférieure ou égale à 34	20	1,00	100

2.2.6 Le diagramme « stem-and-leaf »

Un diagramme « stem-and-leaf » (diagramme « branche et feuille ») est une représentation graphique qui révèle simultanément l'ordre et la forme d'un ensemble de données. Pour illustrer l'utilisation d'un diagramme « stem-and-leaf », considérons l'ensemble de données du tableau 2.8. Ces données sont les résultats d'un test d'aptitude comprenant 150 questions, effectué par 50 individus ayant récemment passé un entretien pour un poste chez Haskens Manufacturing. Les données indiquent le nombre de réponses correctes (fichier en ligne Test d'aptitude).

Pour construire un diagramme « stem-and-leaf », on ordonne les premiers chiffres de chaque observation à gauche d'une ligne verticale. À droite de cette ligne verticale, on rapporte le dernier chiffre de chaque observation. En utilisant la première ligne de données du tableau 2.8 (112, 72, 69, 97 et 107), les premiers pas dans la construction du diagramme « stem-and-leaf » sont les suivants :

6	9
7	2
8	
9	7
10	7
11	2
12	
13	
14	

Par exemple, l'observation 112 est composée du premier chiffre 11 placé à gauche de la ligne et du chiffre 2 placé à droite. De manière similaire, l'observation 72 est composée du chiffre 7, placé à gauche de la ligne et du chiffre 2, placé à droite. En continuant

Tableau 2.8 ***Nombre de réponses correctes au test d'aptitude***

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

à placer le dernier chiffre de chaque observation sur la ligne correspondant à ses premiers chiffres, on obtient :

6	9	8								
7	2	3	6	3	6	5				
8	6	2	3	1	1	0	4	5		
9	7	2	2	6	2	1	5	8	8	5	4
10	7	4	8	0	2	6	6	0	6		
11	2	8	5	9	3	5	9				
12	6	8	7	4							
13	2	4									
14	1										

Avec cette organisation des données, ordonner les chiffres de chaque ligne de la plus petite à la plus grande valeur est simple. On obtient ainsi le diagramme « stem-and-leaf » présenté ci-dessous.

6	8	9								
7	2	3	3	5	6	6				
8	0	1	1	2	3	4	5	6		
9	1	2	2	2	4	5	5	6	7	8	8
10	0	0	2	4	6	6	6	7	8		
11	2	3	5	5	8	9	9				
12	4	6	7	8							
13	2	4									
14	1										

Les nombres à gauche de la ligne verticale (6, 7, 8, 9, 10, 11, 12, 13 et 14) forment la « branche » et chaque chiffre à droite de la ligne verticale correspond à une « feuille ». Par exemple, considérons la première ligne ayant pour branche le chiffre 6 et pour feuilles les chiffres 8 et 9.

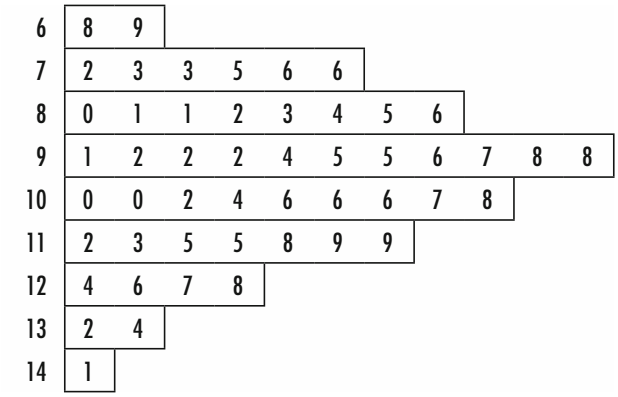
6	8	9

La signification de cette ligne est que deux observations ont pour premier chiffre le 6 : 68 et 69. De même, la seconde ligne

7	2	3	3	5	6	6

indique que six observations ont pour premier chiffre le 7 : 72, 73, 73, 75, 76 et 76.

Pour se concentrer sur la forme du diagramme, traçons un rectangle contenant les feuilles de chaque branche. Nous obtenons la représentation suivante.



En effectuant une rotation à 90° dans le sens inverse des aiguilles d'une montre, on obtient une représentation des données similaire à un histogramme avec les classes 60-69, 70-79, 80-89, etc.

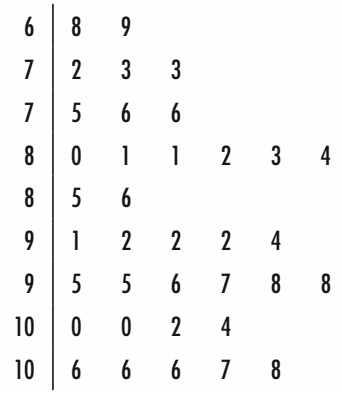
Bien que le diagramme « stem-and-leaf » semble fournir la même information qu'un histogramme, il présente deux avantages supplémentaires.

1. Le diagramme « stem-and-leaf » est plus facile à construire à main levée.

2. À l'intérieur d'une classe, le diagramme « stem-and-leaf » fournit plus d'informations que l'histogramme, puisqu'il donne la valeur des observations.

De la même manière qu'une distribution de fréquence ou un histogramme n'ont pas un nombre absolu de classes, le diagramme « stem-and-leaf » n'a pas un nombre absolu de lignes ou de branches. Si on pense que le diagramme original condense trop les données, on peut facilement étendre le diagramme en utilisant deux ou plusieurs branches pour chaque premier(s) chiffre(s). Par exemple, pour utiliser deux lignes pour chaque premier(s) chiffre(s), on place toutes les observations se terminant par le chiffre 0, 1, 2, 3 ou 4 sur une ligne et toutes les observations se terminant par le chiffre 5, 6, 7, 8 ou 9 sur une seconde ligne. Le diagramme « stem-and-leaf » élargi ci-dessous illustre ces propos.

Dans un diagramme « stem-and-leaf » élargi, quand une valeur de branche est notée deux fois, à la première valeur de la branche sont associées les valeurs des feuilles comprises entre 0 et 4 et à la seconde, les valeurs des feuilles comprises entre 5 et 9.



11	2 3
11	5 8 9 9
12	4
12	6 7 8
13	2 4
13	
14	1

Notez que les observations 72, 73 et 73, dont la feuille a une valeur comprise entre 0 et 4, sont regroupées sur la première branche de valeur 7. Les observations 75, 76 et 76, dont la feuille a une valeur comprise entre 5 et 9, sont regroupées sur la deuxième branche de valeur 7. Ce diagramme « stem-and-leaf » élargi est similaire à une distribution de fréquence dont les intervalles seraient 65-69, 70-74, 75-79, etc.

L'exemple précédent illustre le cas d'un diagramme « stem-and-leaf » pour des données ayant au plus trois chiffres. Les diagrammes « stem-and-leaf » pour des données ayant plus de trois chiffres sont possibles. Par exemple, considérons les données suivantes sur le nombre de hamburgers vendus dans un fast-food, par semaine, pendant 15 semaines.

1565	1852	1644	1766	1888	1912	2044	1812
1790	1679	2008	1852	1967	1954	1733	

Le diagramme « stem-and-leaf » pour ces données est représenté ci-dessous.

Unité de la feuille = 10

15	6
16	4 7
17	3 6 9
18	1 5 5 8
19	1 5 6
20	0 4

Un seul chiffre est utilisé pour définir chaque feuille dans un diagramme « stem-and-leaf ». L'unité de la feuille indique par combien multiplier les nombres du diagramme pour approcher les données initiales. L'unité de la feuille peut être égale à 100, 10, 1 ou 0,1.

Notez qu'un seul chiffre est utilisé pour constituer chaque feuille et que les trois premiers chiffres de chaque observation ont été utilisés pour constituer la branche. En haut du diagramme, nous avons spécifié l'unité de la feuille, égale à 10. Pour illustrer l'interprétation des valeurs du diagramme, considérons la première branche, 15, et la feuille qui lui est associée, 6. En les combinant, on obtient le nombre 156. Pour approcher les observations originales, on doit multiplier ce nombre par 10, l'unité de la feuille. Ainsi, $156 \times 10 = 1560$ est une approximation de l'observation originale, utilisée pour construire le diagramme « stem-and-leaf ». Bien qu'il ne soit pas possible de reconstruire les données exactes à partir du diagramme « stem-and-leaf », la convention qui consiste à utiliser

un seul chiffre pour chaque feuille permet de construire des diagrammes « stem-and-leaf » pour des données comportant un grand nombre de chiffres. Lorsque l'unité de la feuille n'est pas précisée, elle est supposée égale à 1.

REMARQUES

1. Un diagramme en barres et un histogramme sont fondamentalement deux choses identiques. Tous deux sont une représentation graphique des données exprimées sous forme d'une distribution de fréquence. Un histogramme est simplement un diagramme en barres sans séparation entre les rectangles. Pour certaines données quantitatives discrètes, une séparation entre les rectangles est toutefois appropriée. Considérez, par exemple, le nombre de cours qu'un étudiant suit. Les données ne peuvent être que des nombres entiers. Des valeurs intermédiaires telles que 1,5 ou 2,73 ne sont pas possibles. Par contre, avec des données quantitatives continues, telles que les données sur la durée des audits du tableau 2.4, une séparation entre les rectangles n'est pas appropriée.

2. Les valeurs adéquates des limites de classe pour des données quantitatives dépendent du niveau de précision des données. Par exemple, pour les données sur la durée des audits du tableau 2.4, les valeurs des limites de classe étaient des nombres entiers puisque les données avaient été arrondies au jour le plus proche. Si les données avaient été arrondies au dixième de jour le plus proche (par exemple, 12,3, 14,4, etc.), alors les limites auraient été établies en dixième de jour. Par exemple, les limites de la première classe auraient été 10,0-14,9. Si les données avaient été arrondies au centième de jour le plus proche (par exemple, 12,34, 14,45, etc.), alors les limites auraient été établies en centième de jour. Par exemple, les limites de la première classe auraient été 10,00-14,99.

3. Une *classe ouverte* est une classe qui a seulement une limite inférieure ou supérieure. Par exemple, supposez que dans l'exemple sur la durée des audits du tableau 2.4, deux des audits aient nécessité 58 et 65 jours. Plutôt que de continuer la liste des intervalles de 5 jours avec les classes 35-39, 40-44, 45-49, etc., on peut simplifier la distribution de fréquence en considérant une classe ouverte « 35 et plus ». Cette classe aurait une fréquence égale à 2. Le plus souvent, les classes ouvertes apparaissent à la fin de la distribution. Parfois, une classe ouverte apparaît au début de la distribution et occasionnellement, de telles classes apparaissent aux deux extrémités de la distribution.

4. La dernière valeur d'une distribution de fréquence cumulée est toujours égale au nombre total d'observations. La dernière valeur d'une distribution de fréquence cumulée relative est toujours égale à 1 et celle d'une distribution de fréquence cumulée en pourcentage à 100.

EXERCICES

Méthode

11. Considérer les données suivantes (fichier en ligne Fréquence) :

14	21	23	21	16
19	22	25	16	16
24	24	25	19	16
19	18	19	21	12
16	17	18	23	25
20	23	16	20	19
24	26	15	22	24
20	22	24	22	20

a) Développer une distribution de fréquence en utilisant les classes 12-14, 15-17, 18-20, 21-23 et 24-26.

b) Développer une distribution de fréquence relative et une distribution de fréquence en pourcentage en utilisant les mêmes classes.

12. Considérer la distribution de fréquence suivante.

Classe	Fréquence
10-19	10
20-29	14
30-39	17
40-49	7
50-59	2

Construire les distributions de fréquence cumulée absolue et relative.

13. Construire un histogramme à partir des données de l'exercice 12.

14. Considérer les données suivantes :

8,9	10,2	11,5	7,8	10,0	12,2	13,5	14,1	10,0	12,2
6,8	9,5	11,5	11,2	14,9	7,5	10,0	6,0	15,8	11,5

a) Construire un diagramme de points.

b) Construire une distribution de fréquence.

c) Construire une distribution de fréquence en pourcentage.

15. Construire un diagramme « stem-and-leaf » pour les données suivantes.

11,3	9,6	10,4	7,5	8,3	10,5	10,0
9,3	8,1	7,7	7,5	8,4	6,3	8,8

16. Construire un diagramme « stem-and-leaf » pour les données suivantes. Utiliser une unité de feuille égale à 10.

1161	1206	1478	1300	1604	1725	1361	1422
1221	1378	1623	1426	1557	1730	1706	1689

Applications

17. Le personnel d'un cabinet médical a étudié les temps d'attente des patients qui arrivent au cabinet pour une urgence. Les données suivantes ont été collectées au cours d'un mois (les temps d'attente sont exprimés en minutes).

2 5 10 124 4 5 17 11 8 9 8 12 21 6 8 7 13 18 3

Utiliser les classes 0-4, 5-9, etc.

- **a)** Construire la distribution de fréquence.
- **b)** Construire la distribution de fréquence relative.
- **c)** Construire la distribution de fréquence cumulée.
- **d)** Construire la distribution de fréquence cumulée relative.
- **e)** Quelle est la proportion de patients qui viennent en urgence et qui ont un temps d'attente inférieur ou égal à 9 minutes ?

18. CBSSports.com a développé un système de notation des joueurs de l'Association nationale de basketball (NBA), basé sur plusieurs statistiques de jeu offensif et défensif. Les données suivantes (fichier en ligne PointsJoueursNBA) indiquent le nombre moyen de points gagnés par jeu (PPJ) par les 50 meilleurs joueurs sur une partie de la saison 2012-2013 (site Internet de CBSSports.com, 25 février 2013).

27,0	28,8	26,4	27,1	22,9	28,4	19,2	21,0	20,8	17,6
21,1	19,2	21,2	15,5	17,2	16,7	17,6	18,5	18,3	18,3
23,3	16,4	18,9	16,5	17,0	11,7	15,7	18,0	17,7	14,6
15,7	17,2	18,2	17,5	13,6	16,3	16,2	13,6	17,1	16,7
17,0	17,3	17,5	14,0	16,9	16,3	15,1	12,3	18,7	14,6

Utilisez les classes 10-11,9, 12-13,9, 14-15,9, etc. pour répondre aux questions suivantes :

- **a)** Construire la distribution de fréquence.
- **b)** Construire la distribution de fréquence relative.
- **c)** Construire la distribution de fréquence en pourcentage cumulée.
- **d)** Construire un histogramme pour le nombre moyen de points gagnés par jeu.
- **e)** Les données semblent-elles biaisées ? Expliquer.
- **f)** Quel pourcentage de joueurs marquent en moyenne au moins 20 points par jeu ?

19. Sur la base des quantités de marchandises traitées (en millions de tonnes) sur une année, les ports listés ci-dessous (fichier en ligne Ports) sont les 25 ports les plus actifs des États-Unis (*The 2013 World Almanac*).

Ports

Port	Tonnage (millions de tonnes)	Port	Tonnage (millions de tonnes)
Baltimore	39,6	Norfolk Harbor	41,6
Baton Rouge	55,5	Pascagoula	37,3
Beaumont	77,0	Philadelphie	34,0
Corpus Christi	73,7	Pittsburgh	33,8
Duluth-Superior	36,6	Plaquemines	55,8
Houston	227,1	Port Arthur	30,2
Hunnington	61,5	Savannah	34,7
Lake Charles	54,6	Louisiane du Sud	236,3
Long Beach	75,4	Saint Louis	30,8
Los Angeles	62,4	Tampa	34,2
Mobile	55,7	Texas City	56,6
La Nouvelle Orléans	72,4	Valdez	31,9
New York	139,2		

a) Quel est le tonnage traité le plus élevé ? Quel est le tonnage traité le plus faible ?

b) Utiliser une largeur de classe de 25 pour construire une distribution de fréquence de ces données, en commençant avec 25-49,9, 50-74,9, 75-99,9, etc.

c) Construire un histogramme. Interpréter l'histogramme.

20. La London School of Economics et la Harvard Business School ont étudié le déroulement d'une journée d'un président directeur général (PDG). L'étude a montré que les PDG passaient en moyenne 18 heures par semaine en réunion, durée qui n'inclut pas les conférences téléphoniques, les repas d'affaires et les évènements publics (*The Wall Street Journal*, 14 février 2012). Sont repris ci-dessous le temps passé en réunion, par semaine (en heures) pour un échantillon de 25 PDG.

14	15	18	23	15
19	20	13	15	23
23	21	15	20	21
16	15	18	18	19
19	22	23	21	12

a) Quelle est la durée minimale passée en réunion par semaine ? La durée maximale ?

b) Utiliser une largeur de classe de 2 heures pour construire des distributions de fréquence absolue et en pourcentage de ces données.

c) Construire un histogramme. Commenter la forme de la distribution.

Grandes sociétés

21. *Fortune* établit une liste des plus importantes sociétés américaines en termes de chiffre d'affaires annuel. Le tableau suivant (fichier en ligne Grandes sociétés) indique le chiffre d'affaires annuel des 50 plus importantes sociétés, exprimé en milliards de dollars (site Internet de *CNN Money*, 15 janvier 2010).

Société	Chiffre d'affaires	Société	Chiffre d'affaires
Amerisource Bergen	71	Lowe's	48
Archer Daniels Midland	70	Marathon Oil	74
AT&T	124	McKesson	102
Bank of America	113	Medco Health	51
Berkshire Hathaway	108	MetLife	55
Boeing	61	Microsoft	60
Cardinal Health	91	Morgan Stanley	62
Caterpillar	51	Pepsico	43
Chevron	263	Pfizer	48
Citigroup	112	Procter & Gamble	84
ConocoPhilips	231	Safeway	44
Costco Wholesale	72	Sears Holdings	47
CVS Caremark	87	State Farm Insurance	61
Dell	61	Sunoco	52
Dow Chemical	58	Target	65
Exxon Mobil	443	Time Warner	47
Ford Motors	146	United Parcel Service	51
General Electric	149	United Technologies	59
Goldman Sachs	54	United Health Group	118
Hewlett-Packard	118	Valero Energy	118
Home Depot	71	Verizon	97
IBM	104	Walgreen	59
JP Morgan Chase	101	Walmart	406
Johnson & Johnson	64	WellPoint	61
Kroger	76	Wells Fargo	52

a) Construire une distribution de fréquence (classes 0-49, 50-99, 100-149, etc.).

b) Construire une distribution de fréquence relative.

c) Construire une distribution de fréquence cumulée.

d) Construire une distribution de fréquence cumulée relative.

e) Que vous apprennent ces distributions de fréquence sur le chiffre d'affaires annuel des plus grandes sociétés américaines.

f) Construire un histogramme. Commenter la forme de la distribution.

g) Quelle est la plus importante société américaine et quel est son chiffre d'affaires annuel ?

22. Le magazine *Entrepreneur* classe les franchises selon des indices de performance comme le taux de croissance, le nombre de points de vente, les coûts d'installation et la stabilité financière. Le nombre de points de vente des 20 plus importantes franchises aux États-Unis (fichier en ligne Franchise) est fourni ci-dessous (*The World Almanac*, 2012).

Franchise

Franchise	Nombre de points de vente aux États-Unis	Franchise	Nombre de points de vente aux États-Unis
Hampton Inns	1 864	Jan-Pro Franchising Intl. Inc.	12 394
ampm	3 183	Hardee's	1 901
McDonald's	32 805	Pizza Hut Inc.	13 281
7-Eleven Inc.	37 496	Kumon Math & Reading Centers	25 199
Supercuts	2 130	Dunkin' Donuts	9 947
Days Inn	1 877	KFC Corp.	16 224
Vanguard Cleaning Systems	2 155	Jazzercice Inc.	7 683
Servpro	1 572	Anytime Fitness	1 618
Subway	34 871	Matco Tools	1 431
Denny's Inc.	1 668	Stratus Building Solutions	5 018

Utiliser les classes de 0 à 4 999, de 5 000 à 9 999, de 10 000 à 14 999, etc., pour répondre aux questions suivantes.

a) Construire une distribution de fréquence absolue et en pourcentage du nombre de points de vente aux États-Unis pour ces franchises.

b) Construire un histogramme à partir de ces données.

c) Commenter la forme de la distribution.

23. Le rapport Nielsen sur la technologie à la maison fournit des informations sur la technologie domestique et son usage. Les données suivantes correspondent aux heures d'utilisation d'un ordinateur au cours d'une semaine par un échantillon de 50 personnes (fichier en ligne Ordinateur).

Ordinateur

4,1	1,5	10,4	5,9	3,4	5,7	1,6	6,1	3,0	3,7
3,1	4,8	2,0	14,8	5,4	4,2	3,9	4,1	11,1	3,5
4,1	4,1	8,8	5,6	4,3	3,3	7,1	10,3	6,2	7,6
10,8	2,8	9,5	12,9	12,1	0,7	4,0	9,2	4,4	5,7
7,2	6,1	5,7	5,9	4,7	3,9	3,7	3,1	6,1	3,1

Résumer les données en construisant :

a) Une distribution de fréquence (en utilisant une largeur de classe de 3 heures).

b) Une distribution de fréquence relative.

c) Un histogramme.

d) Commenter les résultats quant à l'usage d'un ordinateur à la maison.

24. Le magazine *Money* a listé les métiers qui sont plaisants, bien payés et pérennes dans les 10 années à venir (*Money*, novembre 2009). Le tableau suivant recense les 20 meilleurs métiers, ainsi que le salaire médian et le salaire le plus élevé pour les salariés ayant entre deux et sept années d'expérience. Les données sont exprimées en milliers de dollars (fichier en ligne Métier).

Métier	Salaire médian	Salaire le plus élevé
Chef comptable	81	157
Expert-comptable	74	138
Consultant en protection informatique	100	138
Directeur de la communication	78	135
Analyste financier	80	109
Directeur financier	121	214
Analyste en recherche financière	66	155
Responsable général dans l'hôtellerie	77	146
Responsable des ressources humaines	72	111
Banquier d'affaires	106	221
Analyste des systèmes d'information	83	119
Responsable projet des systèmes d'information	99	140
Responsable marketing	77	126
Responsable qualité	80	122
Représentant	67	125
Auditeur interne sénior	76	106
Développeur de logiciels	79	116
Responsable informatique	110	152
Ingénieur systèmes	87	130
Technicien	67	100

Développer un diagramme « stem-and-lef » à la fois pour le salaire médian et pour le salaire le plus élevé. Quelles informations obtenez-vous sur les salaires de ces métiers ?

25. Un psychologue a développé un nouveau test d'intelligence pour adulte. Les résultats du test effectué par 20 individus sont présentés ci-dessous.

114	99	131	124	117	102	106	127	119	115
98	104	144	151	132	106	125	122	118	118

Construire un diagramme « stem-and-leaf » pour ces données.

26. Le semi-marathon Flying Pig de Cincinnati en 2011 (13,1 miles) a compté 10 897 finalistes (site Internet du Marathon Flying Pig de Cincinnati). Les données suivantes indiquent l'âge d'un échantillon de 40 semi-marathoniens (fichier en ligne Marathon).

49	33	40	37	56
44	46	57	55	32
50	52	43	64	40
46	24	30	37	43
31	43	50	36	61
27	44	35	31	43
52	43	66	31	50
72	26	59	21	47

a) Construire un diagramme « stem-and-leaf » étendu.

b) Quel est le groupe d'âge rassemblant le plus grand nombre de coureurs ?

c) Quel est l'âge le plus fréquent ?

2.3 RÉSUMER DES DONNÉES RELATIVES À DEUX VARIABLES SOUS FORME DE TABLEAUX

Jusqu'ici dans ce chapitre, nous nous sommes concentrés sur les méthodes graphiques et sous forme de tableaux utilisées pour résumer les données d'une variable à un moment précis. Souvent, un dirigeant a besoin de résumer les données relatives à deux variables dans le but de révéler la relation – s'il y en a une – entre ces variables. Dans cette section, nous montrons comment résumer sous forme de tableaux les données relatives à deux variables.

2.3.1 Tabulations croisées

La **tabulation croisée** est un résumé sous forme de tableau des données relatives à deux variables. Bien que les deux variables puissent être qualitatives ou quantitatives, les tabulations croisées dans lesquelles l'une des variables est qualitative et l'autre quantitative sont les plus fréquentes. Nous illustrons ce dernier cas de figure en considérant l'application suivante, fondée sur des données issues de l'enquête sur les restaurants menée par Zagat. Des données sur la qualité et le prix des repas ont été collectées auprès d'un échantillon de 300 restaurants situés dans la région de Los Angeles. Le tableau 2.9 présente les données pour les dix premiers restaurants de l'échantillon. Le niveau de qualité est une variable qualitative qui peut prendre les valeurs bon, très

Tableau 2.9 ***Niveau de qualité et prix des repas de 300 restaurants de Los Angeles***

Restaurant	Niveau de qualité	Prix du repas ($)
1	Bon	18
2	Très bon	22
3	Bon	28
4	Excellent	38
5	Très bon	33
6	Bon	28
7	Très bon	19
8	Très bon	11
9	Très bon	23
10	Bon	13
…	…	…

bon ou excellent. Le prix des repas est une variable quantitative qui varie entre 10 et 49 dollars.

Une tabulation croisée de ces données est présentée dans le tableau 2.10. Dans les marges du tableau sont spécifiées les classes des deux variables. À gauche du tableau, apparaissent en ligne les trois classes de la variable qualité (bon, très bon, excellent). En haut du tableau, apparaissent en colonne les quatre classes de la variable prix (10-19 $, 20-29 $, 30-39 $ et 40-49 $). Pour chaque restaurant de l'échantillon, on a un niveau de qualité et le prix du repas. Ainsi, chaque restaurant de l'échantillon est associé à une cellule de la tabulation croisée, à l'intersection de l'une des lignes et de l'une des colonnes. Par exemple, le restaurant numéro 5 est réputé de très bonne qualité et pratique un prix égal à 33 dollars. Ce restaurant est donc comptabilisé dans la cellule située à l'intersection de la colonne 3 et de la ligne 2 du tableau 2.10. Pour construire un tableau de tabulation croisée, on comptabilise simplement le nombre de restaurants qui appartiennent à chacune des cellules du tableau.

> Le fait de grouper les données d'une variable quantitative nous permet de traiter la variable quantitative comme s'il s'agissait d'une variable qualitative lors de la création d'une tabulation croisée.

Bien que quatre classes de tarif aient été utilisées pour construire la tabulation croisée présentée dans le tableau 2.10, elle aurait pu être effectuée en utilisant un nombre supérieur ou inférieur de classes pour la variable prix du repas. Les considérations à prendre en compte pour décider comment regrouper les données d'une variable quantitative dans une tabulation croisée sont identiques à celles qui président au choix du nombre de classes à utiliser lorsque l'on construit une distribution de fréquence pour une variable quantitative. Dans le cadre de cet exemple, quatre classes de tarif ont été jugées être un nombre raisonnable pour révéler une éventuelle relation entre la qualité et le prix du repas.

En examinant le tableau 2.10, on s'aperçoit que le plus grand nombre de restaurants de l'échantillon (64) ont une très bonne qualité et le prix de leurs repas est compris entre 20 et 29 dollars. Seuls deux restaurants sont d'excellente qualité et pratiquent un tarif compris entre 10 et 19 dollars. On peut interpréter de la même façon les autres

Tableau 2.10 ***Tabulation croisée de la qualité et du prix d'un repas dans 300 restaurants de Los Angeles***

	Prix du repas				
Niveau de qualité	**10-19 $**	**20-29 $**	**30-39 $**	**40-49 $**	**Total**
Bon	42	40	2	0	84
Très bon	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

fréquences. De plus, notez que la dernière ligne et la dernière colonne du tableau de tabulation croisée fournissent les distributions de fréquence pour la qualité et le prix des repas séparément. D'après la distribution de fréquence de droite, 84 restaurants sont réputés de bonne qualité, 150 de très bonne qualité et 66 ont une excellente réputation. De la même façon, la dernière ligne en bas du tableau dévoile la distribution de fréquence du prix des repas.

En divisant le total de chaque ligne de la colonne de droite du tableau de tabulation croisée par le total de cette colonne, on obtient les distributions de fréquence relative et en pourcentage pour la variable « qualité ».

Niveau de qualité	Fréquence relative	Fréquence en pourcentage
Bon	0,28	28
Très bon	0,50	50
Excellent	0,22	22
Total	1,00	100

Selon la distribution de fréquence en pourcentage, 28 % des restaurants de l'échantillon sont de bonne qualité, 50 % de très bonne qualité et 22 % d'excellente qualité.

En divisant le total de chaque colonne de la dernière ligne du tableau de tabulation croisée par le total de cette ligne, on obtient les distributions de fréquence relative et en pourcentage pour la variable « prix ».

Prix du repas	Fréquence relative	Fréquence en pourcentage
10-19 $	0,26	26
20-29 $	0,39	39
30-39 $	0,25	25
40-49 $	0,09	9
Total	1,00	100

Notez que la somme des fréquences relatives et en pourcentage ne correspond pas exactement au total (respectivement 1 et 100) du fait des arrondis. Selon la distribution de fréquence en pourcentage, 26 % des repas ont un prix compris entre 10 et 19 dollars, 39 % entre 20 et 29 dollars, etc.

Les distributions de fréquence absolue et relative construites à partir des marges du tableau de tabulation croisée nous fournissent des informations sur chacune des variables individuellement, mais n'apportent aucune information relative à leurs relations. L'intérêt principal d'une tabulation croisée réside dans l'information qu'elle fournit à propos de la relation entre les variables. D'après les résultats du tableau 2.10, il semble que plus les prix sont élevés, meilleure est la qualité du restaurant, et plus les prix sont bas, moins la qualité est bonne.

En convertissant les entrées du tableau en pourcentage, on peut obtenir des informations supplémentaires sur la relation entre les variables. Par exemple, le tableau 2.11 correspond aux fréquences du tableau 2.10 divisées par le total de la ligne considérée et

Tableau 2.11 ***Pourcentages en ligne pour chaque niveau de qualité***

	Prix du repas				
Niveau de qualité	**10-19 $**	**20-29 $**	**30-39 $**	**40-49 $**	**Total**
Bon	50,0	47,6	2,4	0,0	100
Très bon	22,7	42,7	30,6	4,0	100
Excellent	3,0	21,2	42,4	33,4	100

exprimées en pourcentage. Chaque ligne du tableau 2.11 correspond à une distribution de fréquence en pourcentage du prix du repas pour l'un des niveaux de qualité. Pour les restaurants ayant le niveau de qualité le plus faible (bon), on voit que les pourcentages les plus importants sont associés aux restaurants les moins chers (50 % ont des prix variant entre 10 et 19 dollars et 47,6 % ont des prix variant entre 20 et 29 dollars). Pour les restaurants ayant le niveau de qualité le plus élevé (excellent), on voit que les plus importants pourcentages sont associés aux restaurants les plus chers (42,4 % ont des prix variant entre 30 et 39 dollars et 33,4 % ont des prix variant entre 40 et 49 dollars). Ainsi, la même relation entre le prix et la qualité du repas apparaît encore : les repas les plus chers sont associés aux restaurants ayant les niveaux de qualité les plus élevés.

La tabulation croisée est fréquemment utilisée pour examiner la relation entre deux variables. En pratique, les rapports de beaucoup d'études statistiques contiennent un grand nombre de tableaux de tabulation croisée. Dans l'enquête sur les restaurants de Los Angeles, la tabulation croisée est basée sur une variable qualitative (le niveau de qualité) et une variable quantitative (le prix du repas). Des tabulations croisées peuvent également être effectuées lorsque les deux variables sont qualitatives ou quantitatives. Toutefois, lorsque des variables quantitatives sont utilisées, il est nécessaire de regrouper les valeurs que peut prendre la variable dans des classes. Par exemple, dans le cas des restaurants, nous avons regroupé les prix des repas en quatre classes (10-19$, 20-29$, 30-39$, 40-49$).

2.3.2 Le paradoxe de Simpson

Les données de deux ou plusieurs tabulations croisées sont souvent combinées ou agrégées pour produire un résumé montrant comment deux variables sont liées. Dans de tels cas, il convient d'être prudent dans l'interprétation des relations entre deux variables que l'on pourrait faire à partir de la tabulation croisée agrégée. Dans certains cas, les conclusions basées sur la tabulation croisée agrégée peuvent fournir des résultats en contradiction avec les conclusions tirées des données non agrégées. C'est ce que l'on appelle le paradoxe de Simpson. Pour illustrer ce paradoxe, prenons l'exemple de verdicts rendus par deux juges de deux juridictions différentes.

Les juges Ron Luckett et Denis Kendall ont officié à la Cour des plaids communs et au Tribunal municipal au cours des trois dernières années. Certains de leurs jugements étaient renvoyés en appel. Dans la plupart des cas, la Cour d'Appel confirmait

les jugements initiaux, mais parfois, leurs jugements étaient annulés. Pour chaque juge, une tabulation croisée fut développée à partir de deux variables : le jugement en Cour d'Appel (maintenu ou annulé) et le type de juridiction (Cour des plaids communs ou Tribunal municipal). Supposons que les deux tabulations croisées soient ensuite combinées en agrégeant les données concernant le type de juridiction. La tabulation croisée agrégée contient donc deux variables : le jugement en Cour d'Appel (maintenu ou annulé) et le juge (Luckett ou Kendall). Cette tabulation croisée fournit le nombre de jugements en appel pour lesquels le jugement a été maintenu et le nombre de jugements en appel pour lesquels le verdict a été annulé pour les deux juges. La tabulation croisée fournit les résultats suivants, les pourcentages des colonnes apparaissant entre parenthèses à côté de chaque valeur.

	Juge		
Jugement	**Luckett**	**Kendall**	**Total**
Maintenu	129 (86 %)	110 (88 %)	239
Annulé	21 (14 %)	15 (12 %)	36
Total (%)	150 (100 %)	125 (100 %)	275

D'après les pourcentages en colonne, 86 % des jugements prononcés par le juge Luckett ont été confirmés, alors que 88 % des jugements prononcés par le juge Kendall l'ont été. Ainsi, on pourrait conclure que le juge Kendall est plus efficace, un pourcentage plus important de ses jugements étant maintenus en appel.

Les tabulations croisées suivantes présentent séparément les cas jugés par Luckett et Kendall dans les deux juridictions ; les pourcentages des colonnes sont également indiqués entre parenthèses après chaque valeur.

	Juge Luckett		
Jugement	**Tribunal municipal**	**Cour des plaids communs**	**Total**
Maintenu	29 (91 %)	100 (85 %)	139
Annulé	8 (9 %)	18 (15 %)	21
Total (%)	32 (100 %)	118 (100 %)	150

	Juge Kendall		
Jugement	**Tribunal municipal**	**Cour des plaids communs**	**Total**
Maintenu	90 (90 %)	20 (80 %)	110
Annulé	10 (10 %)	5 (20 %)	15
Total (%)	100 (100 %)	25 (100 %)	125

Selon le tableau de tabulation croisée du juge Luckett, ses jugements sont maintenus en appel dans 91 % des cas jugés au Tribunal municipal et dans 85 % des cas jugés à la Cour des plaids communs. Selon le tableau de tabulation croisée du juge Kendall, ses jugements sont maintenus en appel dans 90 % des cas jugés au Tribunal municipal et dans 80 % des cas jugés à la Cour des plaids communs. En comparant les pourcentages des colonnes des tableaux de tabulation croisée, nous constatons que le juge Luckett obtient un meilleur score que le juge Kendall dans les deux juridictions. Ce résultat contredit la conclusion à laquelle nous étions parvenus en agrégeant les données des deux juridictions. Cet exemple illustre le paradoxe de Simpson.

La tabulation croisée initiale était obtenue en agrégeant les données des deux juridictions. Notez que pour les deux juges, le pourcentage d'annulation en appel est plus

important pour les cas jugés à la Cour des plaids communs qu'au Tribunal municipal. Puisque le juge Luckett a jugé un nombre plus important de cas à la Cour des plaids communs, l'agrégation des données est favorable au juge Kendall. Lorsque l'on regarde les tabulations croisées pour les deux juridictions séparément, le juge Luckett apparaît cependant plus performant. Ainsi, dans la tabulation croisée initiale, le *type de juridiction* est une variable cachée qui ne peut être ignorée lorsque l'on cherche à évaluer l'efficacité des deux juges.

À cause du paradoxe de Simpson, il convient d'être extrêmement vigilant lorsque l'on tire des conclusions à partir de données agrégées. Avant de conclure, vous devez chercher à savoir si la forme agrégée ou désagrégée de la tabulation croisée a un impact sur les conclusions de l'étude. Notamment lorsque la tabulation croisée est réalisée à partir de données agrégées, vous devez vous assurer que des variables cachées n'affectent pas les résultats, conduisant à des conclusions différentes lorsque des tabulations croisées agrégées et désagrégées sont effectuées.

EXERCICES

Méthode

27. Les données relatives à 30 observations de deux variables qualitatives x et y sont présentées ci-dessous. Les catégories pour x sont A, B et C ; les catégories pour y sont 1 et 2 (fichier en ligne Tabulation croisée).

Observation	x	y	Observation	x	y
1	A	1	16	B	2
2	B	1	17	C	1
3	B	1	18	B	1
4	C	2	19	C	1
5	B	1	20	B	1
6	C	2	21	C	2
7	B	1	22	B	1
8	C	2	23	C	2
9	A	1	24	A	1
10	B	1	25	B	1
11	A	1	26	C	2
12	B	1	27	C	2
13	C	2	28	A	1
14	C	2	29	B	1
15	C	2	30	B	2

a) Effectuer une tabulation croisée pour les données en utilisant x en ligne et y en colonne.

b) Calculer les pourcentages en ligne.

c) Calculer les pourcentages en colonne.

d) Quelle est la relation, s'il en existe une, entre x et y ?

28. Le tableau ci-dessous présente 20 observations de deux variables quantitatives, x et y (fichier en ligne Tabulation croisée 2).

Observation	x	y	Observation	x	y
1	28	72	11	13	98
2	17	99	12	84	21
3	52	58	13	59	32
4	79	34	14	17	81
5	37	60	15	70	34
6	71	22	16	47	64
7	37	77	17	35	68
8	27	85	18	62	67
9	64	45	19	30	39
10	53	47	20	43	28

a) Effectuer une tabulation croisée pour les données en utilisant x en ligne et y en colonne.

b) Calculer les pourcentages en ligne.

c) Calculer les pourcentages en colonne.

d) Quelle est la relation, s'il en existe une, entre x et y ?

Applications

29. La Daytona 500 est une course automobile sur 500 miles qui a lieu chaque année sur le circuit international de Daytona Beach en Floride. La tabulation croisée suivante indique la marque de la voiture en fonction de la vitesse moyenne des 25 vainqueurs entre 1998 et 2012 (*The 2013 World Almanac*).

	Vitesse moyenne en miles par heure					
Marque	**130-139,9**	**140-149,9**	**150-159,9**	**160-169,9**	**170-179,9**	**Total**
Buick	1					1
Chevrolet	3	5	4	3	1	16
Dodge		2				2
Ford	2	1	2	1		6
Total	6	8	6	4	1	25

a) Calculer les pourcentages en ligne.

b) Quel pourcentage de vainqueurs conduisant une Chevrolet a gagné avec une vitesse moyenne d'au moins 150 miles par heure ?

c) Calculer les pourcentages en colonne.

d) Quel pourcentage de vainqueurs conduisant à une vitesse moyenne comprise entre 160 et 169,9 miles par heure conduisait une Chevrolet ?

30. La tabulation croisée suivante indique la vitesse moyenne des 25 vainqueurs selon les années de la course automobile Daytona 500 (*The 2013 World Almanac*).

	Année					
Vitesse moyenne	**1988-1992**	**1993-1997**	**1998-2002**	**2003-2007**	**2008-2012**	**Total**
130-139,9	1			2	3	6
140-149,9	2	2	1	2	1	8
150-159,9		3	1	1	1	6
160-169,9	2		2			4
170-179,9			1			1
Total	5	5	5	5	5	25

a) Calculer les pourcentages en ligne.

b) Quelle est la relation apparente entre la vitesse moyenne des vainqueurs et l'année ? Qu'est-ce qui peut expliquer cette relation ?

31. Récemment, la direction du golf Oak Tree a reçu quelques plaintes concernant les conditions du parcours de golf. Plusieurs joueurs se plaignaient de la trop grande rapidité du parcours. Plutôt que de réagir sur la seule base de ces réclamations, la direction du golf a mené une enquête auprès de 100 joueurs et 100 joueuses. Les résultats de l'enquête sont résumés ci-dessous.

Hommes

	Conditions du parcours	
Handicap	Trop rapides	Parfaites
Moins de 15	10	40
15 ou plus	25	25

Femmes

	Conditions du parcours	
Handicap	Trop rapides	Parfaites
Moins de 15	1	9
15 ou plus	39	51

a) Combiner ces deux tabulations croisées en une seule avec, en ligne, le sexe des joueurs (homme ou femme) et en colonne, les conditions de parcours (trop rapides, parfaites). Dans quel groupe, le pourcentage de joueurs trouvant le parcours trop rapide est-il le plus élevé ?

b) Référez-vous aux tabulations croisées initiales. Pour les joueurs avec un faible handicap (les meilleurs), quel groupe (homme ou femme) considère le parcours comme trop rapide ?

c) Référez-vous aux tabulations croisées initiales. Pour les joueurs avec un fort handicap, quel groupe (homme ou femme) considère le parcours comme trop rapide ?

d) Quelles conclusions pouvez-vous tirer des préférences des hommes et des femmes concernant la vitesse du parcours ? Les conclusions tirées en (a) sont-elles cohérentes avec celles tirées des questions (b) et (c) ? Expliquer les incohérences apparentes.

32. Le tableau 2.12 fournit des informations relatives à 45 fonds mutuels qui font partie du *Morningstar Funds 500*, en 2008 (fichier en ligne Fonds mutuels). L'ensemble de données inclut les cinq variables suivantes :

- Le type de fonds : domestique (D), international (I) ou à revenu fixe (F)
- La valeur nette de l'actif (en dollars) : le prix de clôture de l'action
- Le rendement moyen sur cinq ans (%) : le rendement annuel moyen du fonds au cours des cinq dernières années
- Le ratio de dépenses (%) : le pourcentage des actifs déduit chaque année fiscale pour couvrir les frais de gestion du fonds
- Le classement Morningstar : le classement (en nombre d'étoiles) ajusté du risque de chaque fonds ; l'échelle Morningstar va de 1 à 5 étoiles.

a) Préparer une tabulation croisée des données sur le type de fonds (en ligne) et le rendement annuel moyen au cours des cinq dernières années (en colonne). Utiliser les classes 0-9,99, 10-19,99, 20-29,99, 30-39,99, 40-49,99 et 50-59,99 pour le rendement moyen sur cinq ans.

b) Construire la distribution de fréquence pour les données sur le type de fonds.

c) Construire la distribution de fréquence pour les données sur le rendement moyen à cinq ans.

d) Dans quelle mesure le tableau de tabulation croisée vous a aidé à construire les distributions de fréquence des questions (b) et (c) ?

e) Quelles conclusions pouvez-vous tirer à propos du type de fonds et du rendement moyen au cours des 5 dernières années ?

33. En vous référant aux données du tableau 2.12,

a) Préparer une tabulation croisée des données sur le type de fonds (en ligne) et le ratio de dépenses (en colonne). Utiliser les classes 0,25-0,49, 0,50-0,74, 0,75-0,99, 1,00-1,24 et 1,25-1,49 pour le ratio des dépenses.

b) Construire la distribution de fréquence des données relatives au ratio des dépenses.

c) Quelles conclusions pouvez-vous tirer à propos du type de fonds et du ratio de dépenses ?

34. Le fichier en ligne Faillite bancaire contient une liste de 492 banques qui ont fait faillite entre 2000 et 2012 (site Internet de la Federal Deposit Insurance Corporation, 9 mars 2013). Le fichier contient le nom de la banque, la ville, l'État et l'année de la faillite.

a) Construire une tabulation croisée avec l'État en ligne et l'année de la faillite en colonne.

b) Quels sont les trois États dans lesquels les faillites ont été les plus nombreuses ?

c) Donner la distribution de fréquence des faillites bancaires par année. Quelle conclusion pouvez-vous en tirer quant à l'évolution des faillites bancaires au cours du temps ?

35. Le guide relatif aux économies de carburant du département américain à l'énergie fournit des données sur la consommation des voitures et camions (site Internet « Fuel Economy »,

Tableau 2.12 ***Données financières d'un échantillon de 45 fonds mutuels***

Fonds	Type de fonds	Valeur nette de l'actif ($)	Rendement moyen sur 5 ans (%)	Ratio de dépenses (%)	Classement Morningstar
Amer Cent Inc & Growth Inv	D	28,88	12,39	0,67	2 étoiles
American Century International Disc	I	14,37	30,53	1,41	3 étoiles
American Century Tax-Free Bond	F	10,73	3,34	0,49	4 étoiles
American Century Ultra	D	24,94	10,88	0,99	3 étoiles
Ariel	D	46,39	11,32	1,03	2 étoiles
Artisan Inlt Val	I	25,52	24,95	1,23	3 étoiles
Artisan Small Cap	D	16,92	15,67	1,18	3 étoiles
Baron Asset	D	50,67	16,77	1,31	5 étoiles
Brandywine	D	36,58	18,14	1,08	4 étoiles
Brown Cap Small	D	35,73	15,85	1,20	4 étoiles
Buffalo Mid Cap	D	15,29	17,25	1,02	3 étoiles
Delafield	D	24,32	17,77	1,32	4 étoiles
DFA U.S. Micro Cap	D	13,47	17,23	0,53	3 étoiles
Dodge & Cox Income	F	12,51	4,31	0,44	4 étoiles
Fairholme	D	31,86	18,23	1,00	5 étoiles
Fidelity Contrafund	D	73,11	17,99	0,89	5 étoiles
Fidelity Municipal Income	F	12,58	4,41	0,45	5 étoiles
Fidelity Overseas	I	48,39	23,46	0,90	4 étoiles
Fidelity Sel Electronics	D	45,60	13,50	0,89	3 étoiles
Fidelity Sh-Term Bond	F	8,60	2,76	0,45	3 étoiles
Fidelity	D	39,85	14,40	0,56	4 étoiles
FPA New Income	F	10,95	4,63	0,62	3 étoiles
Gabelli Asset AAA	D	49,81	16,70	1,36	4 étoiles
Greenspring	D	23,59	12,46	1,07	3 étoiles
Janus	D	32,26	12,81	0,90	3 étoiles
Janus Worlwide	I	54,83	12,31	0,86	2 étoiles
Kalmar Gr Val Sm Cp	D	15,30	15,31	1,32	3 étoiles
Managers Freemont Bond	F	10,56	5,14	0,60	5 étoiles
Marsico 21st Century	D	17,44	15,16	1,31	5 étoiles
Mathews Pacific Tiger	I	27,86	32,70	1,16	3 étoiles
Meridan Value	D	31,92	15,33	1,08	4 étoiles
Oakmark I	D	40,37	9,51	1,05	2 étoiles
PIMCO Emerg Mkts Bd D	F	10,68	13,57	1,25	3 étoiles
RS Value A	D	26,27	23,68	1,36	4 étoiles
T. Rowe Price Latin America	I	53,89	51,10	1,24	4 étoiles
T. Rowe Price Mid Val	D	22,46	16,91	0,80	4 étoiles
Templeton Growth A	I	24,07	15,91	1,01	3 étoiles
Thornburg Value A	D	37,53	15,46	1,27	4 étoiles

USAA Income	F	12,10	4,31	0,62	3 étoiles
Vanguard Equity-Inc	D	24,42	13,41	0,29	4 étoiles
Vanguard Global Equity	I	23,71	21,77	0,64	5 étoiles
Vanguard GNMA	F	10,37	4,25	0,21	5 étoiles
Vanguard Sht-Tm TE	F	15,68	2,37	0,16	3 étoiles
Vanguard Sm Cp Idx	D	32,58	17,01	0,23	3 étoiles
Wasatch Sm Cp Growth	D	35,41	13,98	1,19	4 étoiles

8 septembre 2012). Une partie des données relatives à 149 voitures de différentes tailles (compactes, moyennes et grandes) est reprise dans le tableau 2.13. L'ensemble de données contient les variables suivantes :

- Taille : Compacte, Moyenne ou Grande
- Motorisation : Taille du moteur en litres
- Cylindrée : Nombre de cylindres dans le moteur
- Roues motrices : Avant (AV), Arrière (AR) ou 4 roues motrices (4)
- Type de carburant : Sans plomb (SP) ou Ordinaire (O)
- Consommation en ville : Consommation urbaine en nombre de miles par gallon
- Consommation sur autoroute : Consommation sur autoroute en miles par gallon

Tableau 2.13 ***Données sur la consommation de carburant pour 311 voitures***

Données carburant 2012

Voiture	Taille	Motorisation	Cylindrée	Roues motrices	Type de carburant	Consommation urbaine	Consommation sur autoroute
1	Compacte	2.0	4	AV	SP	21	30
2	Compacte	2.0	4	4	SP	21	29
3	Compacte	2.0	4	4	SP	21	31
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
94	Moyenne	3,5	6	4	O	17	25
95	Moyenne	2,5	4	AV	O	23	33
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
148	Grande	6,7	12	AR	SP	11	18
149	Grande	6,7	12	AR	SP	11	18

L'ensemble de données complet est contenu dans le fichier en ligne nommé Données Carburant 2012.

a) Préparer une tabulation croisée des données relatives à la taille (en ligne) et à la consommation sur autoroute (en colonne). Utiliser les classes 15-19, 20-24, 25-29, 30-34 et 35-39 pour la consommation sur autoroute.

b) Commenter la relation entre la taille et la consommation sur autoroute.

c) Préparer une tabulation croisée des données relatives au nombre de roues motrices (en ligne) et à la consommation en ville (en colonne). Utiliser les classes 5-9, 10-14, 15-19, 20-24, 25-29, 30-34 et 35-39 pour la consommation en ville.

d) Commenter la relation entre le nombre de roues motrices et la consommation en ville.

e) Préparer une tabulation croisée des données relatives au type de carburant (en ligne) et à la consommation en ville (en colonne). Utiliser les classes 5-9, 10-14, 15-19, 20-24, 25-29, 30-34 et 35-39 pour la consommation en ville.

f) Commenter la relation entre le type de carburant et la consommation en ville.

2.4 RÉSUMER DES DONNÉES RELATIVES À DEUX VARIABLES SOUS FORME DE GRAPHIQUES

Dans la section précédente, nous avons montré comment se servir d'une tabulation croisée pour résumer les données relatives à deux variables et aider à révéler la relation entre ces variables. Dans la plupart des cas, une représentation graphique est plus utile pour appréhender les informations et les tendances contenues dans les données.

Dans cette section, nous introduisons plusieurs représentations graphiques pour explorer les relations entre deux variables. Représenter les données de façon créative peut être très révélateur et nous permet d'en déduire des « inférences de bon sens » basées sur notre capacité à comparer, mettre en exergue et reconnaître des tendances de façon visuelle. Nous commençons avec une discussion sur les nuages de points et les courbes de tendance.

2.4.1 Nuage de points et courbe de tendance

Un **nuage de points** est une représentation graphique de la relation entre deux variables quantitatives et la **tendance** est une droite qui fournit une approximation de la relation. À titre d'illustration, considérons la relation entre les campagnes publicitaires et les ventes d'un magasin d'équipement hi-fi à San Francisco. À dix reprises au cours des trois derniers mois, le magasin a mené une campagne publicitaire télévisée en fin de semaine pour promouvoir ses ventes. Les dirigeants veulent découvrir s'il existe une relation entre le nombre de spots publicitaires diffusés en fin de semaine et les ventes réalisées au cours de la semaine suivante. Le tableau 2.14 contient les données sur les ventes du magasin en milliers de dollars pendant les dix semaines qui ont suivi la diffusion d'un spot publicitaire.

Tableau 2.14 ***Données d'échantillon pour le magasin d'équipement hi-fi***

Semaine	Nombre de spots publicitaires x	Volume des ventes (centaines de dollars) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

La figure 2.7 reproduit le nuage de points et la tendance[1] pour les données du tableau 2.14. Le nombre de spots publicitaires (x) est représenté sur l'axe horizontal, les ventes (y) sur l'axe vertical. Pour la semaine 1, $x = 2$ et $y = 50$. Un point ayant ces coordonnées est dessiné sur le diagramme. Des points similaires sont dessinés pour les neuf autres semaines. Notez que durant deux semaines, un seul spot publicitaire fut diffusé, durant deux autres semaines, deux spots ont été diffusés, etc.

Le nuage de points de la figure 2.7 révèle une relation positive entre le nombre de spots publicitaires diffusés et les ventes réalisées. Un volume de vente plus important est associé à un nombre plus important de spots publicitaires. La relation n'est pas parfaite dans la mesure où tous les points ne sont pas situés sur une même ligne droite. Cependant, la forme générale des points et la tendance suggèrent une relation globalement positive.

La figure 2.8 représente les principales formes des nuages de points et le type de relation qu'elles suggèrent. Le graphique en haut à gauche décrit une relation positive comme celle que nous venons de voir. Le graphique en haut à droite ne révèle aucune relation apparente entre les variables. Le graphique du bas décrit une relation négative, y ayant tendance à décroître quand x augmente.

1 L'équation de la droite de tendance est $y = 36{,}15 + 4{,}95x$. La pente de la droite de tendance est égale à 4,95 et l'ordonnée à l'origine (le point où la droite coupe l'axe des ordonnées) à 36,15. Nous discuterons en détail de l'interprétation de la pente et de l'ordonnée à l'origine pour une droite de tendance linéaire au chapitre 12, lorsque nous étudierons la régression linéaire simple.

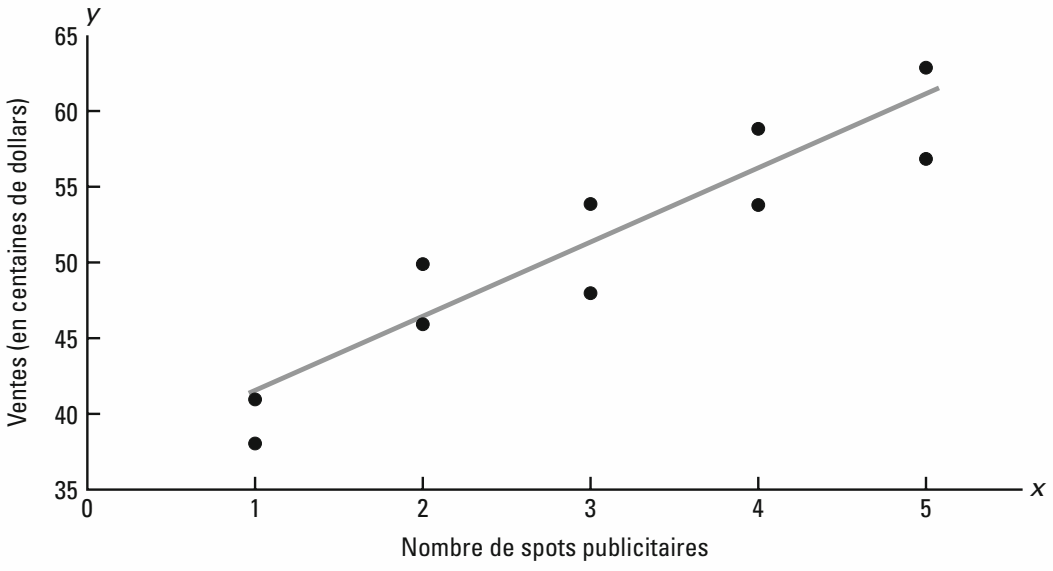


Figure 2.7 ***Nuage de points et droite de tendance pour le magasin de hi-fi***

2.4.2 Diagrammes en barres empilées et côte-à-côte

Dans la section 2.1, nous avons dit qu'un diagramme en barres est une représentation graphique pertinente pour décrire des données qualitatives résumées par une distribution de fréquence absolue, relative ou en pourcentage. Les diagrammes en barres empilées ou côte-à-côte sont des extensions des diagrammes en barres classiques utiles pour représenter et comparer deux variables. En représentant deux variables sur un même graphique, nous pouvons mieux appréhender la relation qui existe entre ces variables.

Un **diagramme en barres côte-à-côte** est une représentation graphique pour décrire sur un même graphique plusieurs diagrammes. Pour illustrer la construction d'un diagramme côte-à-côte, nous reprenons l'exemple relatif aux données sur la qualité et le prix des repas d'un échantillon de 300 restaurants situés dans la région de Los Angeles. La qualité du repas est une variable qualitative qui peut prendre les valeurs Bon, Très bon et Excellent. Le prix du repas est une variable quantitative dont la valeur est comprise entre 10 et 49 dollars. La tabulation croisée figurant dans le tableau 2.10 indique que les données relatives au prix du repas ont été regroupées en quatre classes : 10-19 dollars, 20-29 dollars, 30-39 dollars et 40-49 dollars. Nous utiliserons ces classes pour construire le diagramme en barres côte-à-côte.

La figure 2.9 représente le diagramme côte-à-côte obtenu à partir de ces données. La couleur de chaque barre indique le niveau de qualité (noir = bon, gris foncé = très bon et gris clair = excellent). La hauteur de chaque barre correspond à la fréquence à laquelle ce niveau de qualité est observé pour chaque catégorie de prix. Placer côte-à-côte la fréquence à laquelle une qualité donnée est observée pour chaque catégorie de

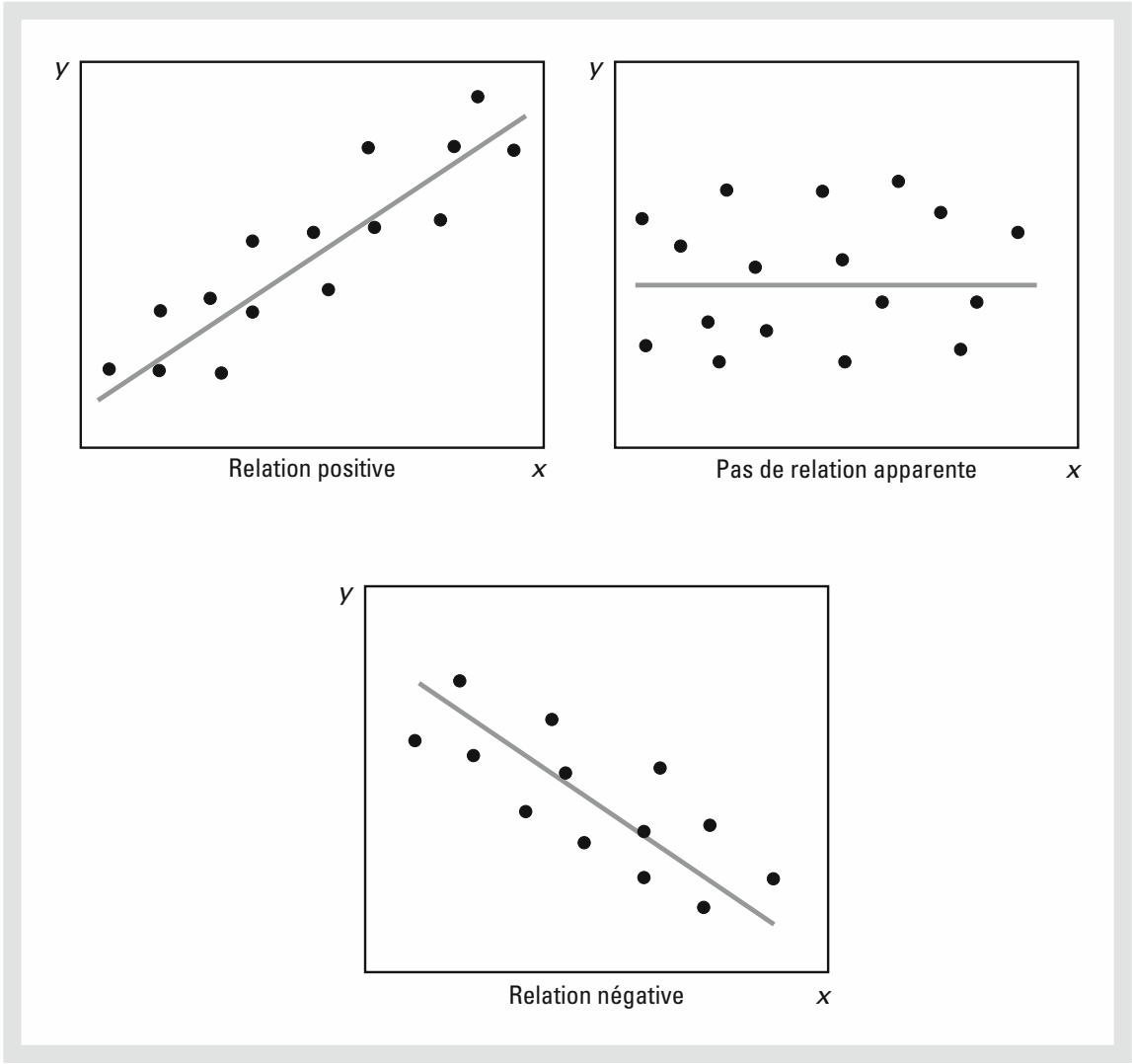


Figure 2.8 ***Types de relations décrites par des nuages de points***

prix nous permet de déterminer rapidement la qualité d'une catégorie de prix particulière. Nous voyons que les repas appartenant à la catégorie de prix la plus faible (10-19 dollars) sont les plus fréquemment considérés comme bon ou très bon mais rarement comme excellent. Les repas appartenant à la catégorie de prix la plus élevée (40-49 dollars) offrent une image différente. La plupart du temps, les repas entrant dans cette catégorie de prix sont considérés comme excellents ; certains comme très bons mais aucun n'est considéré comme « seulement » bon.

La figure 2.9 fournit également des indications sur la relation entre le prix et la qualité d'un repas. Notez que lorsque le prix augmente (lorsque l'on se dirige de la gauche vers la droite du graphique), la hauteur des barres noires a tendance à diminuer et la hauteur des barres de couleur gris clair à augmenter. Cela indique que lorsque les prix augmentent, la note attribuée aux repas a tendance à s'améliorer. La note très bon, comme

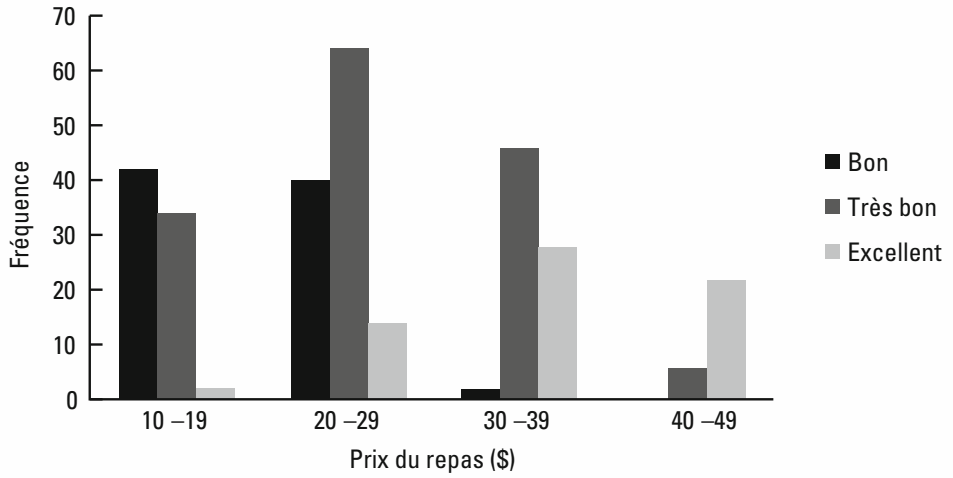


Figure 2.9 ***Diagramme en barres côte-à-côte pour les données sur la qualité et le prix des repas***

on s'y attend, tend à être plus fréquente dans les classes de prix intermédiaires comme le révèle la dominance des barres de couleur gris foncé dans le milieu du graphique.

Les diagrammes en barres empilées sont un autre moyen de représenter et de comparer deux variables sur le même graphique. Un diagramme en barres empilées est un graphique en barres dans lequel chaque barre est segmentée en rectangle de couleur différentes représentant la fréquence relative de chaque classe de façon similaire à un diagramme circulaire. Pour illustrer un diagramme en barres empilées, nous utilisons les données sur la qualité et le prix des repas résumées dans le tableau de tabulation croisée (tableau 2.10).

Nous pouvons convertir les données de fréquence du tableau 2.10 en pourcentage par colonne en divisant chaque élément d'une colonne donnée par le total de cette colonne Par exemple, 42 des 78 restaurants dont le prix est compris entre 10 et 19 dollars sont réputés « bon ». Le tableau 2.15 fournit les pourcentages en colonne pour chaque catégorie de prix. En utilisant les données du tableau 2.15, nous avons construit le diagramme en barres empilées de la figure 2.10. Dans la mesure où le diagramme en barres empilées est basé sur des pourcentages, la figure 2.10 indique encore plus clairement que la figure 2.9 la relation entre les variables. Lorsque l'on passe de la catégorie de prix la plus basse (10-19 dollars) à la plus élevée (40-49 dollars), la longueur des segments noirs diminue et celle des segments gris clairs augmente.

Tableau 2.15 ***Pourcentages en colonne pour chaque catégorie de prix***

Niveau de qualité	Prix du repas			
	10-19 $	20-29 $	30-39 $	40-49 $
Bon	53,8 %	33,9 %	2,6 %	0,0 %
Très bon	43,6	54,2	60,5	21,4
Excellent	2,6	11,9	36,8	78,6
Total	100 %	100 %	100 %	100 %

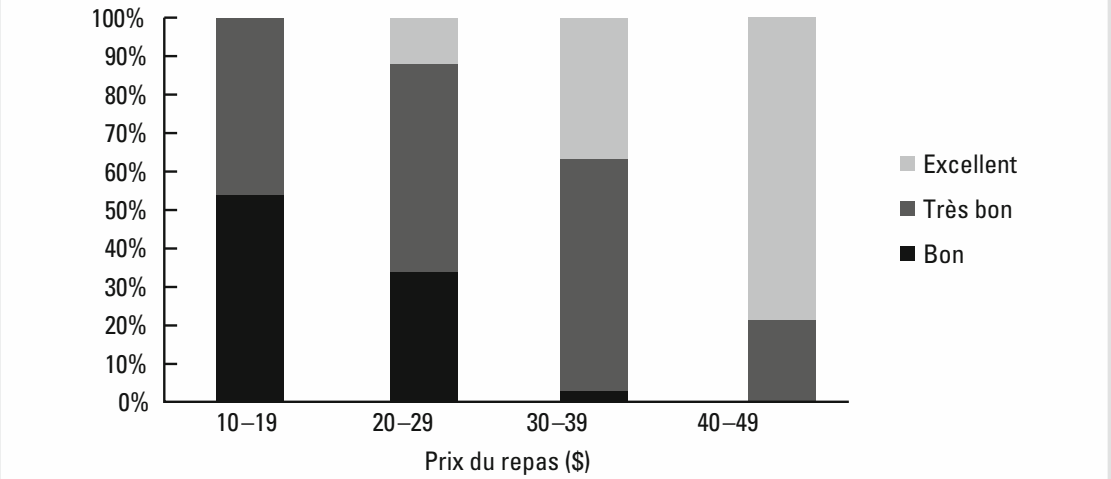


Figure 2.10 ***Diagramme en barres empilées pour les données sur la qualité et le prix des repas***

REMARQUES

Un diagramme en barres empilées peut être utilisé pour représenter des fréquences plutôt que des fréquences en pourcentage. Dans ce cas, les différents segments de couleur de chaque barre représentent la contribution au total de cette barre, plutôt que la contribution en pourcentage.

EXERCICES

Méthode

36. Vingt observations relatives à deux variables quantitatives, x et y, sont fournies ci-dessous (fichier en ligne Nuage de Points).

Observation	x	y	Observation	x	y
1	–22	22	11	–37	48
2	–33	49	12	34	–29
3	2	8	13	9	–18
4	29	–16	14	–33	31
5	–13	10	15	20	–16
6	21	–28	16	–3	14
7	–13	27	17	–15	18
8	–23	35	18	12	17
9	14	–5	19	–20	–11
10	3	–3	20	–7	–22

a) Représenter le nuage de points de la relation entre x et y.

b) Quelle est la relation, si elle existe, entre x et y ?

37. Considérez les données suivantes relatives à deux variables qualitatives. La première variable, x, peut prendre les valeurs A, B, C ou D. La seconde variable, y, peut prendre les valeurs I ou II. Le tableau suivant fournit la fréquence à laquelle chaque combinaison survient.

x	y	
	I	*II*
A	143	857
B	200	800
C	321	679
D	420	580

a) Construire un diagramme en barres côte-à-côte avec x sur l'axe horizontal.

b) Commenter la relation entre x et y.

38. Le tableau de tabulation croisée ci-dessous résume les données relatives à deux variables qualitatives, x et y. La variable x peut prendre les valeurs faible, moyen ou élevé et la variable y peut prendre les valeurs oui ou non.

x	y		Total
	Oui	*Non*	
Faible	20	10	30
Moyen	15	35	50
Élevé	20	5	25
Total	55	50	105

a) Calculer les pourcentages en ligne.

b) Construire un diagramme en barres empilées de la fréquence en pourcentage avec x sur l'axe horizontal.

2.4.3 Applications

39. Une étude sur la vitesse (en miles par heure) et la consommation de carburant (distance en miles parcourue avec un gallon) de voitures de taille moyenne a fourni les données suivantes (fichier en ligne MPG) :

Vitesse	30	50	40	55	30	25	60	25	50	55
Consommation	28	25	25	23	30	32	21	35	26	25

a) Représenter le nuage de points avec la vitesse sur l'axe horizontal et la consommation sur l'axe vertical.

b) Commenter toute relation qui apparaîtrait entre ces deux variables.

40. Le site Internet Current Results fournit la liste des températures minimales et maximales moyennes annuelles (en degré Fahrenheit) et les chutes de neige moyennes annuelles (en pouces) pour 51 grandes villes américaines, relevées au cours de la période 1981-2010. Les données figurent dans le fichier en ligne Neige. Par exemple, la température minimale moyenne enregistrée dans la ville de Columbus dans l'Ohio est de 44 degrés et les chutes moyennes de neige annuelles de 27,5 pouces.

a) Représenter le nuage de point avec la température minimale annuelle moyenne sur l'axe horizontal et les chutes de neige annuelles moyennes sur l'axe vertical.

b) Est-ce qu'une relation apparaît entre ces deux variables ?

c) En vous basant sur le nuage de points, commenter tout point qui vous semble inhabituel.

41. Les gens ne se préoccupent souvent pas de leur cœur avant la quarantaine. Pourtant, des études récentes ont montré qu'une surveillance précoce des facteurs de risque comme la tension pouvait être très bénéfique (*The Wall Street Journal*, 10 janvier 2012). Avoir une tension supérieure à la normale, un état connu sous le terme d'hypertension, est un facteur de risque majeur pouvant entraîner le développement d'une maladie cardiaque. Supposez qu'un grand échantillon d'individus d'âges et de sexes différents soit sélectionné et que la tension de chaque individu soit mesurée pour déterminer s'il est hypertendu. Le tableau suivant fournit le pourcentage des individus hypertendus (fichier en ligne Hypertension).

Âge	Homme	Femme
20-34	11,0 %	9,0 %
35-44	24,0 %	19,0 %
45-54	39,0 %	37,0 %
55-64	57,0 %	56,0 %
65-74	62,0 %	64,0 %
75 et +	73,3 %	79,0%

Hypertension

a) Construire un diagramme en barres côte-à-côte avec l'âge sur l'axe horizontal, le pourcentage d'individus hypertendus sur l'axe vertical et un diagramme en barres côte-à-côte basé sur le sexe.

b) Qu'indiquent les graphiques à propos de l'hypertension et de l'âge ?

c) Commenter les différences en termes de sexe.

42. Les smartphones sont des téléphones mobiles permettant de se connecter à Internet, de prendre des photos, d'écouter de la musique et de regarder des vidéos (Centre de Recherche Pew, Internet & American Life Project, 2011). Les résultats d'enquête présentés ci-dessous indiquent le taux de possession d'un smartphone en fonction de l'âge (fichier en ligne Smartphones).

Âge	Smartphone (%)	Autre téléphone mobile (%)	Pas de téléphone mobile (%)
18-24	49	46	5
25-34	58	35	7
35-44	44	45	11
45-54	28	58	14
55-64	22	59	19
65 et +	11	45	44

Smartphone

a) Construire un diagramme en barres empilées pour représenter les données de l'enquête sur le type de téléphone mobile que les gens possèdent. Utiliser l'âge comme variable sur l'axe horizontal.

b) Commenter la relation entre l'âge et le taux de possession d'un smartphone.

c) Selon vous, les résultats de l'enquête seraient-ils différents si l'enquête était menée en 2021 ?

43. Le responsable de la région Nord-Ouest d'une enseigne d'équipements pour des activités de plein air a mené une enquête pour déterminer comment les responsables de trois magasins utilisaient leur temps. Un résumé des résultats est fourni dans le tableau ci-dessous (fichier en ligne Emploi du temps des responsables).

Emploi du temps des responsables

Magasin	Pourcentage du temps de travail hebdomadaire du responsable passé à			
	Réunion	Rapports	Clients	Inactif
Bend	18	11	52	19
Portland	52	11	24	13
Seattle	32	17	37	14

a) Construire un diagramme en barres empilées avec le magasin sur l'axe horizontal et le pourcentage de temps passé à chaque tâche sur l'axe vertical.

b) Construire un diagramme en barres côte-à-côte pour le pourcentage de temps passé à chaque tâche (avec le magasin sur l'axe horizontal).

c) Quel type de diagramme en barres (empilées ou côte-à-côte) préférez-vous pour visualiser ces données ? Pourquoi ?

2.5 VISUALISATION DES DONNÉES : LES MEILLEURES PRATIQUES POUR CRÉER DES GRAPHIQUES PERTINENTS

La visualisation des données est un terme employé pour décrire l'utilisation de graphiques pour résumer et présenter des informations relatives à un ensemble de données. Le but de la visualisation des données est de fournir de façon aussi claire et efficace que possible les informations clés concernant les données. Dans cette section, nous fournissons quelques indications pour créer un graphique pertinent, choisir le type de graphiques appropriés au regard de l'objectif de l'étude, utiliser des tableaux de bord et nous montrons comment le zoo et le jardin botanique de Cincinnati utilisent les techniques de visualisation des données pour améliorer leur processus de décision.

Tableau 2.16 ***Ventes anticipées effectives par région (en milliers de dollars)***

Anticipées-Effectives

Région	Anticipées	Effectives
Nord-Est	540	447
Nord-Ouest	420	447
Sud-Est	575	556
Sud-Ouest	360	341

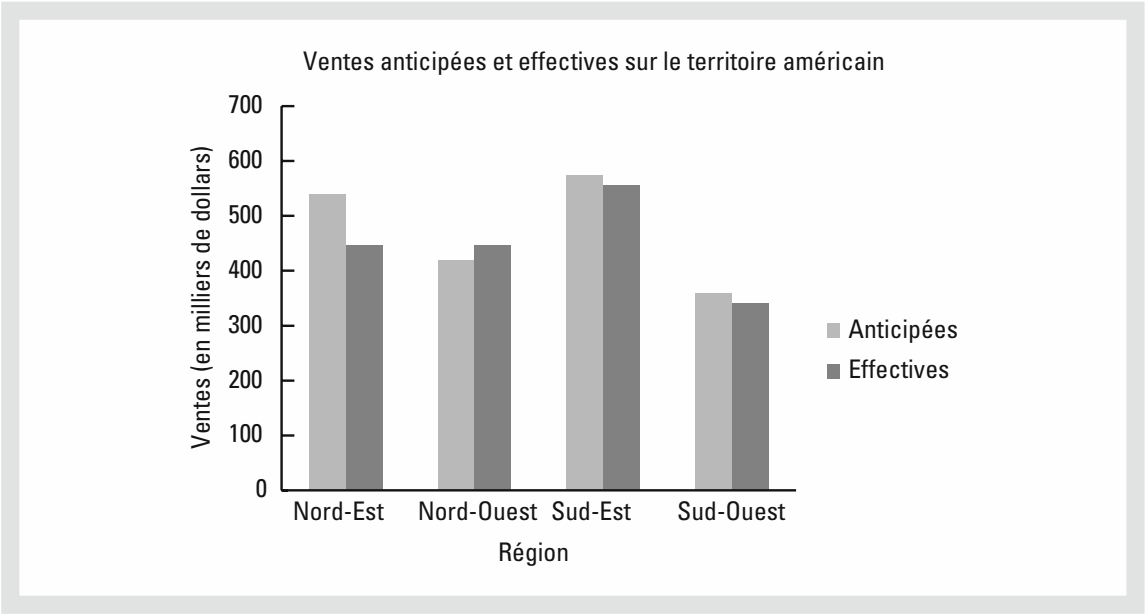


Figure 2.11 ***Diagramme en barres côte-à-côte pour les données sur les ventes anticipées et effectives***

2.5.1 Créer des graphiques pertinents

Les données présentées dans le tableau 2.16 indiquent la valeur des ventes prévisionnelles ou anticipées (en milliers de dollars) et la valeur des ventes effectives ou réalisées (en milliers de dollars) par la société Gustin Chemical l'an passé sur le territoire américain découpé en 4 régions. Notez qu'il y a deux variables quantitatives (les ventes anticipées et les ventes effectives) et une variable qualitative (les régions). Supposez que nous voulions construire un graphique qui permette aux dirigeants de Gustin Chemical de visualiser les ventes effectives de chaque région par rapport aux prévisions et simultanément de visualiser les performances en termes de ventes de chaque région.

Un diagramme en barres côte-à-côte des données sur les ventes anticipées et effectives est représenté sur la figure 2.11. Notez combien ce diagramme en barres permet de comparer facilement les ventes effectives et les ventes anticipées dans une région, ainsi qu'entre les régions. Cette représentation graphique est simple, comporte un titre, est correctement nommée et utilise des couleurs distinctes pour représenter les deux types de données sur les ventes. Remarquez également que l'échelle de l'axe vertical commence à zéro. Les quatre régions sont séparées par un espace de sorte qu'il est clair qu'elles sont distinctes, alors que les ventes anticipées et effectives sont côte-à-côte pour une comparaison simple à l'intérieur de chaque région. Le diagramme en barres côte-à-côte de la figure 2.11 permet de constater facilement que la région Sud-Ouest est celle dans laquelle les ventes à la fois anticipées et réalisées sont les plus faibles et que les ventes réalisées dans la région Nord-Ouest excèdent légèrement les prévisions.

Créer une représentation graphique pertinente relève plus de l'art que de la science. En suivant les indications générales fournies ci-dessous, vous pouvez accroître la probabilité que votre représentation graphique transmette efficacement les informations clés contenues dans les données.

- Nommez de façon claire et concise votre graphique.
- Simplifiez votre graphique. N'utilisez pas trois dimensions lorsque deux sont suffisantes.
- Nommez clairement chaque axe et indiquez les unités de mesure.
- Si des couleurs sont utilisées pour distinguer les catégories, choisissez des couleurs différentes.
- Si plusieurs couleurs ou plusieurs types de rayures sont utilisées, utilisez une légende pour les identifier et placez la légende à côté de la représentation des données.

2.5.2 Choisir le type de graphique

Dans ce chapitre, nous avons présenté un certain nombre de représentations graphiques, dont des diagrammes en barres, des diagrammes circulaires, des diagrammes de points, des histogrammes, des diagrammes stem-and-leaf, des nuages de points, des diagrammes en barres côte-à-côte, des diagrammes en barres empilées. Chacun de ces types de représentation graphique a été développé dans un but précis. Pour fournir des indications quant au choix du type de graphique approprié, nous fournissons maintenant un résumé des types de graphique en fonction de leur finalité. Certaines représentations graphiques peuvent être utilisées de façon appropriée pour atteindre des objectifs différents.

Les graphiques utilisés pour illustrer la distribution des données

- Diagramme en barres – Utilisé pour représenter la distribution de fréquence totale et relative de données qualitatives
- Diagramme circulaire – Utilisé pour représenter la fréquence relative et en pourcentage de données qualitatives
- Diagramme de points – Utilisé pour représenter la distribution de données quantitatives sur l'ensemble des valeurs que prennent les données
- Histogramme – Utilisé pour représenter la distribution de fréquence de données quantitatives sur un ensemble d'intervalles
- Diagramme stem-and-leaf – Utilisé pour montrer à la fois l'ordre et la forme de la distribution de données quantitatives

Les graphiques utilisés pour faire des comparaisons

- Diagramme en barres côte-à-côte – Utilisé pour comparer deux variables
- Diagrammes en barres empilées – Utilisé pour comparer la fréquence relative ou en pourcentage de deux variables qualitatives

Les graphiques utilisés pour révéler des relations

- Le nuage de points – Utilisé pour représenter la relation entre deux variables quantitatives
- La droite de tendance – Utilisée pour approximer la relation entre les données sur un nuage de points

2.5.3 Les tableaux de bord

Les tableaux de bord sont souvent qualifiés de tableaux de bord numériques.

L'un des outils de visualisation des données les plus fréquemment utilisés est le **tableau de bord**. Si vous conduisez une voiture, vous êtes déjà familier avec ce concept de tableau de bord. Dans une voiture, le tableau de bord comporte des gauges et d'autres indicateurs clés pour entretenir le véhicule. Par exemple, les gauges utilisées pour indiquer la vitesse de la voiture, le niveau de carburant, la température du moteur et le niveau d'huile sont essentielles pour assurer la sécurité et la performance de la voiture. Dans certains véhicules, cette information est même visible sur le pare-brise pour fournir une information encore plus efficace au conducteur. Les tableaux de bord de données jouent un rôle similaire dans la prise de décision des dirigeants d'entreprise.

Un tableau de bord est un ensemble de représentations visuelles qui organisent et présentent l'information utilisée pour contrôler la performance d'une entreprise ou d'une organisation de façon simple à lire, comprendre et interpréter. Comme dans le cas d'une voiture dans lequel la vitesse, la réserve de carburant, la température du moteur et le niveau d'huile sont des informations importantes pour conduire de façon efficace, chaque activité économique a des indicateurs de performance clés qui doivent être surveillés pour évaluer la performance d'une entreprise. Parmi ces indicateurs clés, on peut citer les stocks, les ventes journalières, le pourcentage des livraisons réalisées dans le temps imparti et le chiffre d'affaires trimestriel. Un tableau de bord doit fournir un résumé en temps utile (provenant éventuellement de sources différentes) des indicateurs clés de performance qui sont importants pour l'utilisateur et cela, d'une manière informative et agréable.

Pour illustrer l'utilisation d'un tableau de bord dans la prise de décision, nous présentons un exemple relatif à la société Grogan Oil. Grogan a des bureaux situés dans trois villes du Texas : Austin (le siège de la société), Houston et Dallas. Le centre d'appel informatique de la société, qui se trouve dans les bureaux d'Austin, traite les appels des employés qui font face à des problèmes informatiques, relatifs aux logiciels, à Internet ou aux e-mails. Par exemple, si un employé de Dallas a un problème avec un logiciel, l'employé peut appeler le centre d'appel pour obtenir de l'aide.

Le tableau de bord reproduit à la figure 2.12 a été développé pour surveiller la performance du centre d'appel. Ce tableau de bord combine plusieurs graphiques qui permettent de contrôler les indicateurs de performance clés du centre d'appel. Les données présentées concernent l'équipe qui a pris son poste à 8 heures. Le diagramme en barres

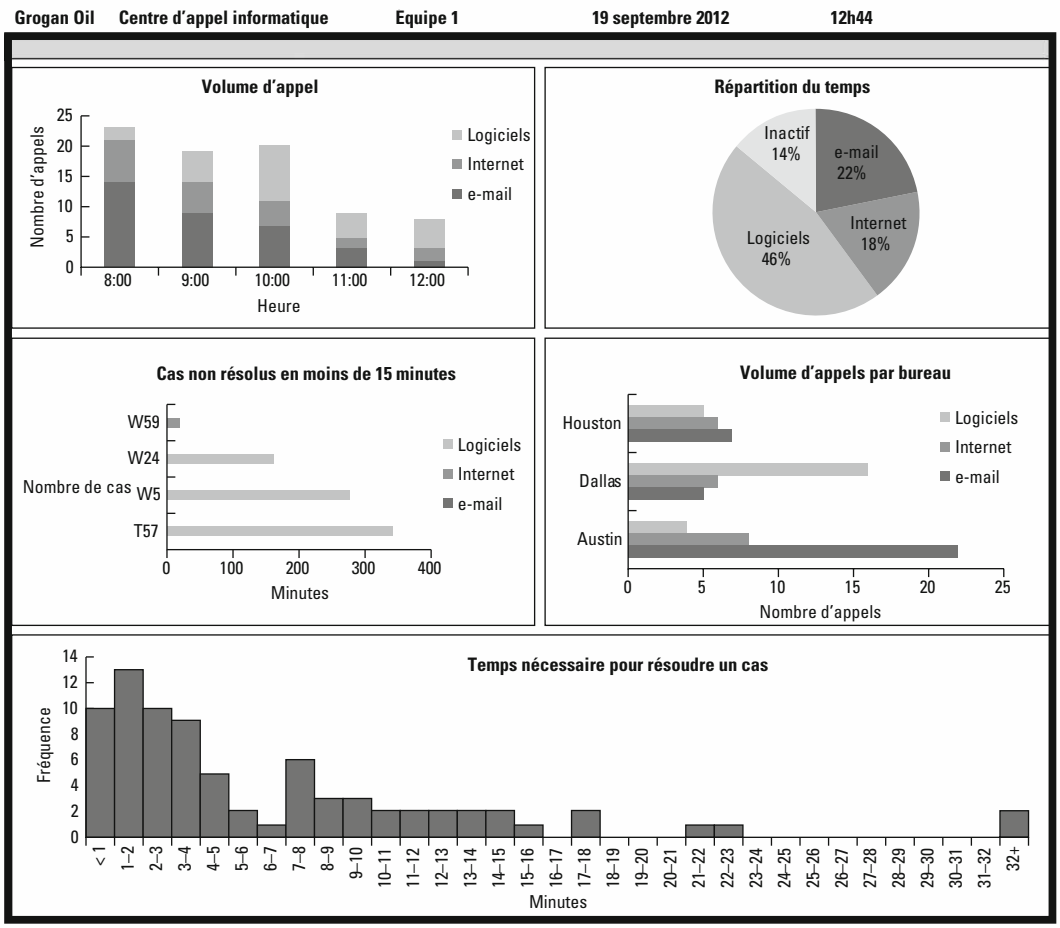


Figure 2.12 *Tableau de bord du centre d'appel informatique de la Grogan Oil*

empilées dans le coin supérieur gauche indique le volume d'appels pour chaque type de problème (logiciels, Internet ou e-mails) par heure. Ce graphique montre que le volume d'appels est plus important durant les premières heures de la journée, les appels concernant des problèmes d'e-mails décroissent au fil des heures et le volume d'appels relatifs aux logiciels est plus important en milieu de matinée. Le diagramme circulaire dans le coin supérieur droit du tableau de bord indique le pourcentage de temps passé par les employés du centre d'appel sur chaque type de problèmes et le temps d'inactivité. Chacun de ces graphiques est utile pour déterminer les besoins en personnel. Par exemple, connaître la raison des appels et le pourcentage d'inactivité peut aider le responsable informatique à s'assurer que suffisamment d'employés ayant le bon niveau d'expertise soient disponibles pour faire face aux besoins.

Le diagramme en barres côte-à-côte situé sous le diagramme circulaire indique le volume d'appels par type de problème pour chacun des bureaux de Grogan. Cela permet au responsable informatique d'identifier rapidement s'il y a un type particulier de

problèmes rencontrés par les employés d'un bureau donné. Par exemple, il apparaît que le bureau d'Austin rencontre un nombre relativement élevé de problèmes d'e-mail. Si la source du problème peut être identifiée rapidement, alors le problème pourra être résolu rapidement. Remarquez également qu'un nombre relativement important de problèmes de logiciel survient dans le bureau de Dallas. Le nombre plus important d'appels dans ce cas était simplement dû au fait que le bureau de Dallas était en train d'installer un nouveau logiciel, et cela a eu pour conséquence d'augmenter le nombre d'appels auprès du centre informatique. Dans la mesure où le responsable informatique avait été alerté par le bureau de Dallas de ce changement la semaine précédente, il avait anticipé l'éventualité d'une augmentation du nombre d'appels en provenance du bureau de Dallas et avait augmenté les ressources en personnel pour traiter ce surplus d'appels attendu.

Le diagramme en barres représenté au milieu, côté gauche, du tableau de bord indique la durée nécessaire pour résoudre chaque cas non résolu en moins de 15 minutes. Ce graphique permet à la société d'identifier rapidement les cas problématiques et de décider d'allouer ou non des ressources additionnelles pour les résoudre. Il a fallu plus de 300 minutes pour résoudre le pire cas, le T57, que l'équipe précédente n'avait pas réussi à solutionner avant sa relève. Pour finir, l'histogramme situé en bas du tableau de bord indique la distribution du temps nécessaire à l'équipe en place pour résoudre les problèmes auxquels elle a été confrontée.

Le tableau de bord de la Grogan Oil illustre l'utilisation d'un tel outil d'un point de vue opérationnel. Le tableau de bord est actualisé en temps réel et utilisé pour prendre des décisions opérationnelles telles que les besoins en personnel. Les tableaux de bord peuvent également être utilisés à des fins tactiques ou stratégiques par les dirigeants. Par exemple, un responsable logistique peut contrôler la performance et le coût de ses sous-traitants. Cela peut l'aider à prendre des décisions quant au mode de transport et au choix des sous-traitants. À un niveau plus élevé, un tableau de bord stratégique peut permettre à la direction d'évaluer rapidement la santé financière de l'entreprise en surveillant des informations financières plus agrégées, le niveau de service et les capacités de production employées.

Les bonnes pratiques en matière de visualisation des données discutées plus haut s'appliquent aux graphiques individuels des tableaux de bord, ainsi qu'au tableau de bord dans son ensemble. En plus de ces bonnes pratiques, il est important de minimiser le besoin de faire défiler l'écran, d'éviter l'usage non nécessaire de couleurs ou de graphiques en trois dimensions et de séparer les graphiques de manière à en améliorer la lecture. Comme pour les graphiques individuels, la simplicité est toujours préférable.

2.5.4 La visualisation des données en pratique : le zoo et le jardin botanique de Cincinnati[2]

Le zoo de Cincinatti, dans l'Ohio, est le second plus ancien zoo au monde. Pour améliorer la prise de décision basée sur les données, la direction a décidé de lier les différentes

2 Les auteurs remercient John Lucas, membre du zoo et du jardin botanique de Cincinnati, de leur avoir fourni cet exemple.

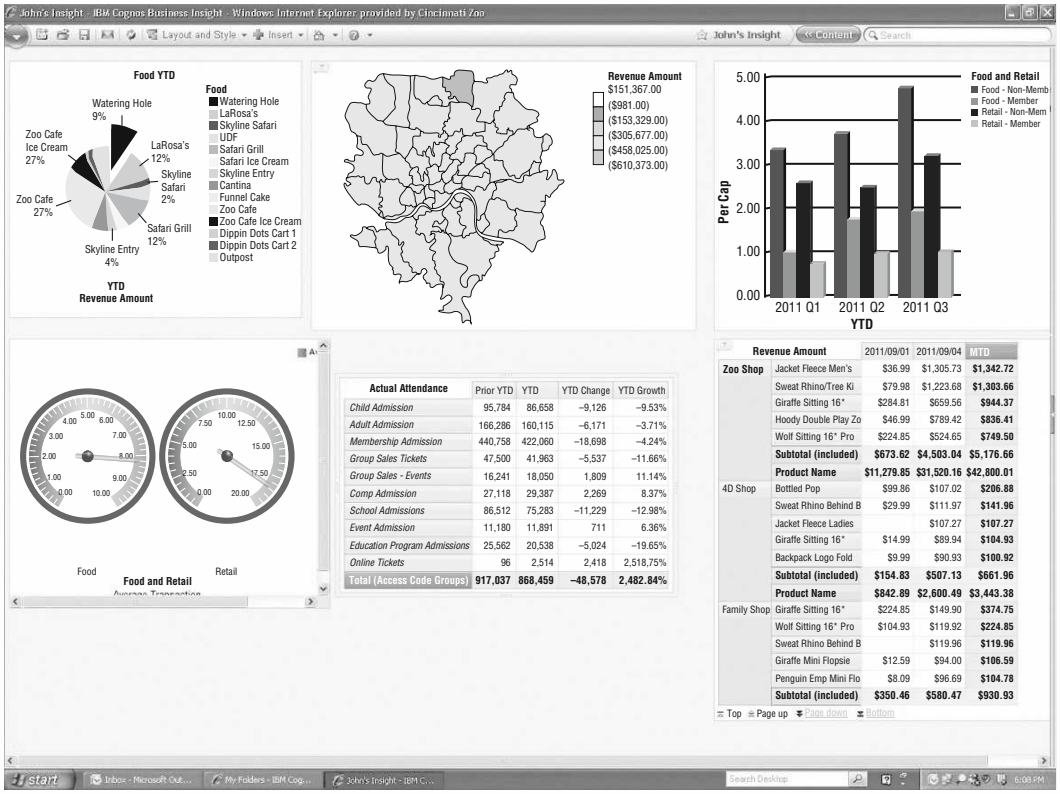


Figure 2.13 ***Le tableau de bord du zoo de Cincinnati***

facettes de son activité et de fournir à des responsables non experts une façon intuitive de mieux comprendre leurs données. Un facteur qui complexifie le problème, est que, les jours d'affluence, les responsables doivent être sur le terrain pour accueillir les visiteurs, vérifier les opérations et anticiper les problèmes qui pourraient survenir. Par conséquent, être en mesure de surveiller ce qui se passe en temps réel était un facteur clé pour décider quoi faire. La direction du zoo en a conclu qu'une stratégie de visualisation des données était nécessaire pour répondre à ce besoin.

Du fait de sa simplicité d'usage, de sa capacité à se réactualiser en temps réel et de sa compatibilité avec les iPad, le zoo de Cincinnati a décidé de déployer la stratégie de visualisation des données offerte par le logiciel Cognos d'IBM. En utilisant ce logiciel, le zoo a conçu le tableau de bord, reproduit à la figure 2.13, pour permettre aux responsables du zoo de surveiller les indicateurs de performance clés suivants :

- Analyse par produit (volume des ventes et valeur des ventes par point de vente à l'intérieur du zoo)
- Analyse géographique (utilisation de cartes et de graphiques pour identifier les endroits où les visiteurs passent leur temps dans le zoo au cours de la journée)

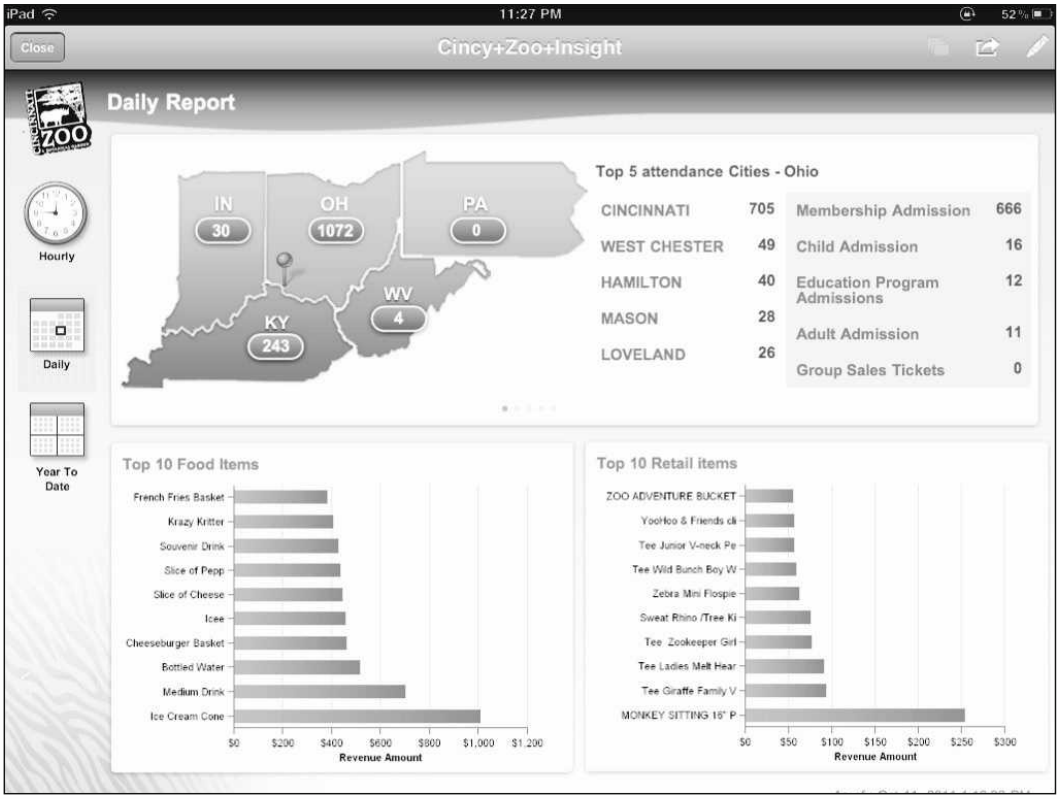


Figure 2.14 ***Le tableau de bord du zoo de Cincinnati***

- Dépenses des clients
- Performance des vendeurs
- Données sur les ventes et les entrées en fonction de la météo
- Performance du programme de fidélité du zoo

Une application mobile pour iPad a également été développée pour permettre aux responsables du zoo d'être à la fois sur le terrain et d'anticiper ce qui se passe en temps réel. Le tableau de bord sur iPad du zoo de Cincinnati, reproduit à la figure 2.14, fournit aux responsables les informations suivantes :

- Les entrées en temps réel, y compris des informations sur les « types » de visiteurs qui entrent dans le zoo
- Des analyses en temps réel sur les produits qui sont vendus
- Une représentation géographique en temps réel des déplacements des visiteurs à l'intérieur du zoo

L'accès aux données présentées sur les figures 2.13 et 2.14 permet aux responsables du zoo de prendre de meilleures décisions quant aux besoins en personnel du zoo,

aux produits qui doivent être stockés en fonction de la météo et d'autres facteurs, et sur la façon de cibler leurs publicités en fonction de données géo-démographiques.

La visualisation des données sur le zoo a eu un impact significatif. Au cours de la première année d'utilisation, le système fut directement responsable d'une augmentation du chiffre d'affaires de plus de 500 000 dollars, d'une fréquentation accrue du zoo, d'une amélioration du service client et d'une réduction des coûts marketing.

REMARQUES

1. Différents logiciels de visualisation des données sont disponibles. Parmi les plus populaires, on trouve Cognos, JMP, Spotfire et Tableau.

2. Les graphiques en radar et en bulle sont deux autres formes de graphiques fréquemment utilisées pour représenter des relations entre plusieurs variables. Cependant, beaucoup d'experts en visualisation des données recommandent de ne pas utiliser ces graphiques en raison de leur complexité. L'usage de représentations graphiques plus simples comme les diagrammes en barres et les nuages de points est recommandé.

3. Un outil très puissant de visualisation des données est le Système d'Information Géographique (SIG). Un SIG se sert de couleurs, de symboles et d'annotations sur une carte pour aider à comprendre comment des variables sont distribuées géographiquement. Par exemple, une société qui cherche à implanter un nouveau centre de distribution peut souhaiter mieux comprendre comment la demande pour son produit varie à travers le pays. Un SIG peut être utilisé pour représenter la demande en identifiant en rouge les régions dans lesquelles la demande est forte, en bleu les régions dans lesquelles la demande est faible et en blanc les régions dans lesquelles le produit n'est pas vendu. Les zones situées près des régions en rouge peuvent s'avérer de bons candidats pour une nouvelle implantation.

RÉSUMÉ

Un ensemble de données, aussi modeste soit sa taille, est souvent difficile à interpréter directement sous sa forme originelle. Des procédures graphiques et sous forme de tableaux permettent d'organiser et de résumer les données, de manière à révéler leur tendance et à les interpréter plus facilement. Les distributions de fréquence absolue, relative ou en pourcentage, les diagrammes en barres et les diagrammes circulaires sont des procédures graphiques et sous forme de tableaux permettant de résumer des données qualitatives. Quand il s'agit de données quantitatives, on peut utiliser les distributions de fréquence absolue, relative ou en pourcentage, les diagrammes de points, les histogrammes, les distributions de fréquence cumulées absolue, relative, en pourcentage, ainsi qu'une technique d'analyse exploratoire des données, le diagramme « stem-and-leaf ».

Pour résumer des données relatives à deux variables, on peut effectuer une tabulation croisée. Le nuage de points est une méthode graphique illustrant la relation entre

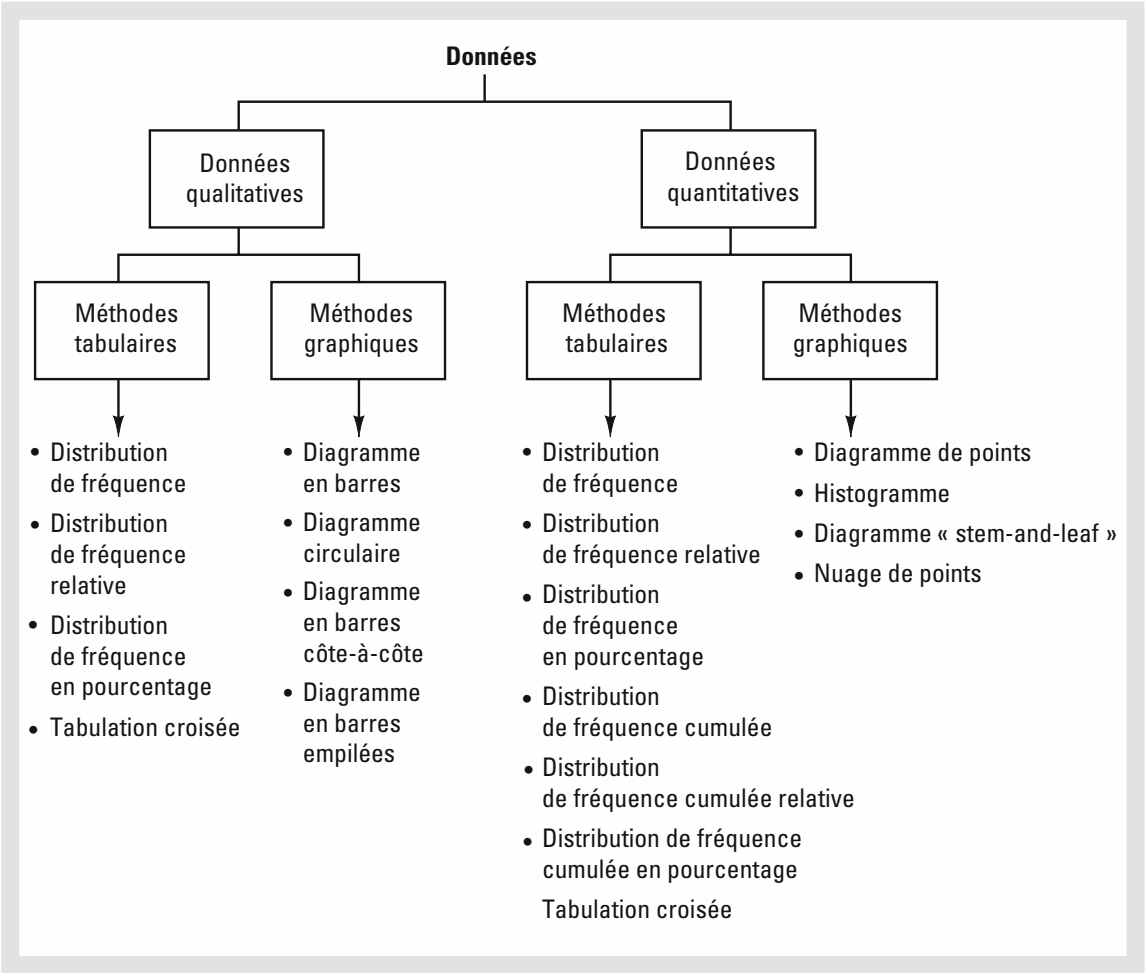


Figure 2.15 ***Le tableau de bord du zoo de Cincinnati***

deux variables quantitatives. Nous avons également montré que les diagrammes en barres côte-à-côte et les diagrammes en barres empilées sont des extensions des diagrammes en barres classiques qui peuvent être utilisées pour représenter et comparer deux variables quantitatives. Des indications pour créer des représentations graphiques pertinentes et choisir le type de graphiques le plus approprié ont été fournies. Les tableaux de bord de données ont été introduits pour illustrer comment un ensemble de représentations visuelles pouvait être développé pour organiser et présenter des informations utiles au contrôle de la performance d'une entreprise de manière simple à lire, comprendre et interpréter. La figure 2.15 résume l'ensemble des méthodes graphiques et sous forme de tableaux présentées dans ce chapitre.

Avec de grands échantillons, les logiciels informatiques sont essentiels pour construire ces résumés graphiques et sous forme de tableaux. Dans les annexes de ce chapitre, nous montrons comment Minitab, Excel et StatTools peuvent être utilisés à cette fin.

GLOSSAIRE

DONNÉES QUALITATIVES Labels ou noms utilisés pour identifier les caractéristiques des observations.

DONNÉES QUANTITATIVES Valeurs numériques qui indiquent des quantités.

VISUALISATION DES DONNÉES Terme utilisé pour décrire l'utilisation de représentations graphiques pour résumer et présenter des informations relatives à un ensemble de données.

DISTRIBUTION DE FRÉQUENCE (ABSOLUE) Résumé des données sous forme d'un tableau, indiquant le nombre (la fréquence) des observations dans chacune des classes.

DISTRIBUTION DE FRÉQUENCE RELATIVE Résumé des données sous forme d'un tableau, indiquant la proportion des observations dans chacune des classes.

DISTRIBUTION DE FRÉQUENCE EN POURCENTAGE Résumé des données sous forme d'un tableau, indiquant le pourcentage des observations dans chacune des classes.

DIAGRAMME EN BARRES Méthode graphique décrivant des données qualitatives résumées sous forme d'une distribution de fréquence absolue, relative ou en pourcentage.

DIAGRAMME CIRCULAIRE Méthode graphique résumant des données, basée sur la subdivision d'un cercle en sections qui correspondent à la fréquence relative pour chaque classe.

CENTRE DE CLASSE Point dans chaque classe qui est à égale distance des limites inférieure et supérieure de la classe.

DIAGRAMME DE POINTS Graphique qui résume des données par le nombre de points placés audessus de chaque valeur de l'ensemble des données représentée sur l'axe horizontal.

HISTOGRAMME Présentation graphique d'une distribution de fréquence absolue, relative ou en pourcentage de données quantitatives, construite en plaçant les classes sur l'axe horizontal et les fréquences absolues, relatives ou en pourcentage sur l'axe vertical.

DISTRIBUTION DE FRÉQUENCE CUMULÉE (ABSOLUE) Résumé sous forme d'un tableau, de données quantitatives indiquant le nombre d'observations dont la valeur est inférieure ou égale à la limite supérieure de chaque classe.

DISTRIBUTION DE FRÉQUENCE CUMULÉE RELATIVE Résumé sous forme d'un tableau, de données quantitatives indiquant la proportion des observations dont la valeur est inférieure ou égale à la limite supérieure de chaque classe.

DISTRIBUTION DE FRÉQUENCE CUMULÉE EN POURCENTAGE Résumé sous forme d'un tableau, de données quantitatives indiquant le pourcentage d'observations dont la valeur est inférieure ou égale à la limite supérieure de chaque classe.

ANALYSE EXPLORATOIRE DE DONNÉES Méthode qui utilise des calculs simples et des graphiques faciles à dessiner pour résumer des données rapidement.

DIAGRAMME « STEM-AND-LEAF » Technique d'analyse exploratoire des données qui, simultanément, ordonne les données quantitatives et fournit des informations sur la forme de la distribution.

TABULATION CROISÉE Résumé sous forme d'un tableau pour deux variables. Les classes de l'une des variables sont notées en ligne ; les classes de l'autre variable sont notées en colonne.

PARADOXE DE SIMPSON Conclusions tirées de deux ou plusieurs tabulations croisées séparément qui se révèlent en contradiction avec celles tirées lorsque les données sont agrégées en une seule tabulation croisée.

NUAGE DE POINTS Illustration graphique de la relation entre deux variables quantitatives.

Une variable est représentée sur l'axe horizontal, l'autre sur l'axe vertical.

Tendance Droite qui fournit une approximation de la relation entre deux variables.

Diagramme en barres côte-à-côte Représentation graphique permettant de décrire des diagrammes en barres multiples sur le même graphique.

Diagramme en barres empilées Diagramme en barres dans lequel chaque barre est séparée en segments rectangulaires de couleurs différentes pour décrire la fréquence relative de chaque classe à la manière d'un diagramme circulaire.

Tableau de bord Ensemble de représentations visuelles qui organisent et présentent des informations utilisées pour contrôler la performance d'une entreprise ou d'une organisation d'une manière simple à lire, comprendre et interpréter.

FORMULES CLÉ

Fréquence relative

$$\frac{\text{Fréquence d'une classe}}{n} \tag{2.1}$$

Largeur approximative d'une classe

$$\frac{\text{Valeur la plus élevée - Valeur la plus faible}}{\text{Nombre de classes}} \tag{2.2}$$

EXERCICES SUPPLÉMENTAIRES

44. Environ 1,5 million de lycéens passent le test d'aptitude scolaire chaque année et près de 80 % des grandes écoles et des universités dans lesquelles l'admission se fait sur dossier, utilisent les résultats à ce test pour décider d'admettre ou non les étudiants (Conseil d'admission, mars 2009). La version actuelle du test d'aptitude comprend trois parties : lecture critique, mathématiques et rédaction. Un score parfait pour les trois parties correspond à 2 400 points. Un échantillon des résultats obtenus au test d'aptitude est présenté ci-dessous (fichier en ligne Résultats test d'aptitude).

1665	1525	1355	1645	1780
1275	2135	1280	1060	1585
1650	1560	1150	1485	1990
1590	1880	1420	1755	1375
1490	1560	940	1390	1175

a) Construire une distribution de fréquence et un histogramme pour ces données. Commencer la première classe avec un résultat de 800 et utiliser une largeur de classe de 200.

b) Discuter de la forme de la distribution.

c) Quelles autres observations peuvent être faites sur les résultats des tests à partir des résumés graphiques et sous forme de tableaux des données.

45. Les Steelers de Pittsburgh ont battu les Cardinals de l'État d'Arizona 27 à 23 lors du 43e Super Bowl. Avec cette victoire, sa sixième en championnat, l'équipe des Steelers de Pittsburg est devenue l'équipe la plus victorieuse dans l'histoire de ce championnat (*Tampa Tribune*, 2 février 2009). Le Super Bowl fut organisé dans huit États différents : Arizona (AZ), Californie (CA), Floride (FL), Géorgie (GA), Louisiane (LA), Michigan (MI), Minnesota (MN) et Texas (TX). Les données présentées dans le tableau suivant indiquent l'État dans lequel les Super Bowl se sont déroulés et le différentiel de points entre l'équipe victorieuse et le perdant (fichier en ligne Super Bowl).

Super Bowl	État	Écart de points	Super Bowl	État	Écart de points	Super Bowl	État	Écart de points
1	CA	25	16	MI	5	31	LA	14
2	FL	19	17	CA	10	32	CA	7
3	FL	9	18	FL	19	33	FL	15
4	LA	16	19	CA	22	34	GA	7
5	FL	3	20	LA	36	35	FL	27
6	FL	21	21	CA	19	36	LA	3
7	CA	7	22	CA	32	37	CA	27
8	TX	17	23	FL	4	38	TX	3
9	LA	10	24	LA	45	39	FL	3
10	FL	4	25	FL	1	40	MI	11
11	CA	18	26	MN	13	41	FL	12
12	LA	17	27	CA	35	42	AZ	3
13	FL	4	28	GA	17	43	FL	4
14	CA	12	29	FL	23			
15	LA	17	30	AZ	10			

a) Construire une distribution de fréquence et un diagramme en barres pour les données sur l'État dans lequel le Super Bowl s'est déroulé.

b) Quelles conclusions pouvez-vous tirer de votre résumé à la question (a) ? Quel est le pourcentage de Super Bowls qui se sont déroulés en Floride ou en Californie ? Quel est le pourcentage de Super Bowls qui se sont déroulés dans les États du Nord ou les États plus froids ?

c) Construire un diagramme « stem-and-leaf » étendu pour l'écart de points entre l'équipe victorieuse et le perdant. Construire un histogramme.

d) Quelles conclusions pouvez-vous tirer des graphiques construits à la question (c) ? Quel est le pourcentage de Super Bowls qui ont été remportés d'une courte victoire, avec un écart de points inférieur à 5 ? Quel est le pourcentage de Super Bowls remportés avec un écart de points supérieur ou égal à 20 ?

e) La victoire la plus courte fut remportée par les Giants de New York contre les Buffalo Bills. Où ce jeu s'est-il déroulé et quel fut l'écart de points ? L'écart de points le plus important dans l'histoire de ce championnat a été observé lorsque les

49e de San Francisco ont battu les Broncos de Denver. Où ce jeu s'est-il déroulé et quel fut l'écart de points ?

46. Des données fournies ci-dessous indiquent la population par État en millions de personnes (*The World Almanac*, 2012, fichier en ligne Population2012).

État	Population	État	Population
Alabama	4,8	Montana	0,9
Alaska	0,7	Nebraska	1,8
Arizona	6,4	Nevada	2,7
Arkansas	2,9	New Hampshire	1,3
Californie	37,3	New Jersey	8,8
Colorado	5,0	Nouveau Mexique	2,0
Connecticut	3,6	New York	19,4
Delaware	0,9	Caroline du Nord	9,5
Floride	18,8	Dakota du Nord	0,7
Géorgie	9,7	Ohio	11,5
Hawaii	1,4	Oklahoma	3,8
Idaho	1,6	Oregon	4,3
Illinois	12,8	Pennsylvanie	12,7
Indiana	6,5	Rhode Island	1,0
Iowa	3,0	Caroline du Sud	4,6
Kansas	2,9	Dakota du Sud	0,8
Kentucky	4,3	Tennessee	6,3
Louisiane	4,5	Texas	25,1
Maine	1,3	Utah	2,8
Maryland	5,8	Vermont	0,6
Massachusetts	6,5	Virginie	8,0
Michigan	9,9	Washington	6,7
Minnesota	5,3	Virginie Occidentale	1,9
Mississippi	3,0	Wisconsin	5,7
Missouri	6,0	Wyoming	0,6

a) Construire des distributions de fréquence absolue et en pourcentage et un histogramme. Utiliser une largeur de classe de 2,5 millions.

b) Discuter de l'asymétrie de la distribution.

c) Quelles observations pouvez-vous faire sur la population des 50 États ?

47. La capacité d'une start-up à lever des fonds est un facteur clé de succès. Les fonds levés (en millions de dollars) par 50 start-up apparaissent ci-dessous (*The World Street Journal*, 10 mars 2011 ; fichier en ligne StartUp).

81	61	103	166	168
80	51	130	77	78
69	119	81	60	20

StartUp

73	50	110	21	60
192	18	54	49	63
91	272	58	54	40
47	24	57	78	78
154	72	38	131	52
48	118	40	49	55
54	112	129	156	31

a) Construire un diagramme « stem-and-leaf ».

b) Commenter ce diagramme.

BBB

48. Des plaintes de consommateurs sont fréquemment enregistrées par le bureau « Better Business ». En 2011, les industries qui ont le plus fait l'objet de plaintes auprès de ce bureau étaient les banques, les compagnies de télévision par câble et satellite, les agences de recouvrement, les fournisseurs de téléphones mobiles et les concessionnaires automobiles (*USA Today*, 16 avril 2012). Les résultats relatifs à un échantillon de 200 plaintes sont contenus dans le fichier en ligne BBB.

a) Indiquer la fréquence et la fréquence en pourcentage de plaintes par industrie.

b) Construire un diagramme en barres de la distribution de fréquence en pourcentage.

c) Quelle industrie a le nombre de plaintes le plus élevé ?

d) Commenter la distribution de fréquence en pourcentage des plaintes.

Tableau 2.17 ***Rendement des dividendes des sociétés composant l'indice Dow Jones industriel***

Société	Rendement des dividendes (%)	Société	Rendement des dividendes (%)
3M	3,6	IBM	2,1
Alcoa	1,3	Intel	3,4
American Express	2,9	Johnson & Johnson	3,6
AT&T	6,6	JPMorgan Chase	0,5
Bank of America	0,4	Kraft Foods	4,4
Boeing	3,8	McDonald's	3,4
Caterpillar	4,7	Merck	5,5
Chevron	3,9	Microsoft	2,5
Cisco Systems	0,0	Pfizer	4,2
Coca-Cola	3,3	Procter & Gamble	3,4
DuPont	5,8	Travelers	3,0
ExxonMobil	2,4	United Technologies	2,9
General Electric	9,2	Verizon	6,3
Hewlett-Packard	0,9	Wal-Mart	2,2
Home Depot	3,9	Walt Disney	1,5

49. Le rendement des dividendes correspond au dividende versé chaque année par une société, exprimé en pourcentage du prix de l'action (dividende divisé par le prix de l'action multiplié par 100). Le rendement des dividendes des sociétés composant l'indice Dow Jones Industriel est fourni dans le tableau 2.17 (*The Wall Street Journal*, 8 juin 2009) et en ligne dans le fichier Rendement des dividendes.

a) Construire des distributions de fréquence absolue et en pourcentage.

b) Construire un histogramme.

c) Discuter de la forme de la distribution.

d) Que vous apprennent les résumés graphiques et sous forme de tableaux sur le rendement des dividendes des sociétés composant l'indice Dow Jones Industriel ?

e) Quelle société présente le rendement le plus élevé ? Si l'action de cette société est actuellement vendue à 14 dollars et que vous achetez 500 actions, quel dividende cet investissement génèrera-t-il en un an ?

50. Le bureau de recensement américain estime les caractéristiques de la population américaine grâce à une enquête que le bureau mène tous les dix ans. Ci-dessous est présentée une tabulation croisée de l'âge et du diplôme le plus élevé obtenu (site Internet du bureau de recensement américain, 9 mars 2013).

Âge	Sans baccalauréat	Niveau baccalauréat	Sans diplôme universitaire	Niveau licence	Niveau maîtrise	Niveau doctorat	Total
25-34	4766	11175	7765	3903	9860	3657	41126
35-44	4732	11568	6593	4166	8858	4530	40447
45-54	4616	14559	7413	4705	8434	4616	44343
55-64	3681	11079	6213	3256	6583	4637	35359
65-74	3563	7418	3290	1383	2955	2326	20935
75 et +	4344	6639	2472	812	2101	1289	17657
Total	25702	62438	33656	18225	38791	21055	199867

a) Calculer les pourcentages en ligne.

b) Calculer les pourcentages en colonne. Comparer les distributions de fréquence en pourcentage pour un niveau maîtrise et un niveau doctorat.

51. L'Université Western n'a plus qu'une place à attribuer dans l'équipe de softball féminine cette année. Les deux finalistes en lice sont Allison Fealey et Emily Janson. L'entraîneur a conclu que les qualités défensives et en termes de vitesse des deux joueuses étaient quasiment identiques et que la décision finale serait prise sur la base du meilleur score moyen de frappes. Les tabulations croisées des performances en termes de frappes de chaque joueuse durant leurs années de lycée, en tant que junior puis sénior, sont reprises ci-dessous.

Résultat	Allison Fealey	
	Junior	Sénior
Frappe	15	75
Pas de frappe	25	175
Total (tentatives de frappe)	40	250

Résultat	Emily Janson	
	Junior	Sénior
Frappe	70	35
Pas de frappe	130	85
Total (tentatives de frappe)	200	120

La moyenne de frappes d'un joueur est calculée en divisant le nombre de frappes d'un joueur par le nombre total de tentatives de frappes. Les moyennes sont exprimées par un nombre décimal arrondi à trois chiffres après la virgule.

a) Calculer la moyenne de frappes de chaque joueuse lors de ses années junior. Calculer ensuite la moyenne de frappes de chaque joueuse dans ses années sénior. Sur la base de cette analyse, quelle joueuse devrait être retenue ? Expliquer.

b) Combiner ou agréger les données des années en tant que junior et sénior dans une seule tabulation croisée.

	Joueuse	
Résultat	**Fealey**	**Janson**
Frappe Pas de frappe		
Total (tentatives de frappe)		

Calculer la moyenne de frappes de chaque joueuse pour les deux années combinées. Sur la base de cette analyse, quelle joueuse devrait être retenue ? Expliquer.

c) Les recommandations que vous avez faites en (a) et en (b) sont-elles cohérentes ? Expliquer les incohérences.

52. Le magazine *Fortune* publie une enquête annuelle des meilleures sociétés dans lesquelles travailler. Les données contenues dans le fichier Fortune Best indiquent le rang, le nom de la société, sa taille et le pourcentage de croissance des emplois à temps complet pour les années à venir d'un échantillon de 98 sociétés (site Internet du magazine *Fortune*, 25 février 2013).

a) Construire une tabulation croisée avec le taux de croissance de l'emploi (%) en ligne et la taille de la société en colonne. Utiliser des classes de -10 à -1, 0-9, 10-19 et ainsi de suite pour le taux de croissance.

b) Indiquer la distribution de fréquence pour le taux de croissance de l'emploi et la distribution de fréquence pour la taille.

c) Utiliser la tabulation croisée développée à la question (a) pour construire une tabulation croisée fournissant les pourcentages en colonne.

d) Utiliser la tabulation croisée développée à la question (a) pour construire une tabulation croisée fournissant les pourcentages en ligne.

e) Commenter la relation entre le taux de croissance des emplois à temps complet et la taille de la société.

Tableau 2.18 ***Données relatives à un échantillon d'écoles et d'universités privées***

Universités

École	Année de création	Frais de scolarité (dollars)	Pourcentage de diplômés
Université américaine	1893	36 697	79
Université Baylor	1845	29 754	70
Université Belmont	1951	23 680	68
…	…	…	…
École Wofford	1854	31 710	82
Université Xavier	1831	29 970	79
Université de Yale	1701	38 300	98

53. Le tableau 2.18 présente une partie des données d'un échantillon de 103 écoles et universités privées. L'ensemble complet de données est contenu dans le fichier en ligne nommé Universités. Les données comprennent le nom de l'école ou de l'université, l'année de création de l'institution, les frais de scolarité (sans pension) au cours des années les plus récentes, et le pourcentage d'étudiants qui ont obtenu leur maîtrise en six ans au plus (*The World Almanac*, 2012).

a) Construire une tabulation croisée avec l'année de création en ligne et les frais de scolarité en colonne. Utiliser des classes commençant à 1600 et finissant à 2000 par saut de 50 pour l'année de création. Pour les frais de scolarité, utiliser des classes commençant à 1 et finissant à 45 000 par saut de 5 000.

b) Calculer les pourcentages en ligne pour la tabulation croisée développée à la question (a).

c) Quelle relation, s'il en existe une, remarquez-vous entre l'année de création et les frais de scolarité ?

54. Référez-vous à l'ensemble de données du tableau 2.18.

a) Construire une tabulation croisée avec l'année de création en ligne et le pourcentage de diplômés en colonne. Utiliser des classes commençant à 1600 et finissant à 2000 par saut de 50 pour l'année de création. Pour le pourcentage de diplômés, utiliser des classes commençant à 35 % et finissant à 100 % par saut de 5 %.

b) Calculer les pourcentages en ligne pour la tabulation croisée développée à la question (a).

c) Commenter la relation, s'il en existe une, entre les variables.

55. Référez-vous à l'ensemble de données du tableau 2.18.

a) Dessiner un nuage de points pour illustrer la relation entre l'année de création et les frais de scolarité.

b) Commenter la relation entre les variables.

56. Référez-vous à l'ensemble de données du tableau 2.18.

a) Dessiner un nuage de points pour illustrer la relation entre les frais de scolarité et le pourcentage de diplômés.

b) Commenter la relation entre les variables.

57. Google a changé sa stratégie en matière d'investissement publicitaire (combien et dans quels médias investir). Le tableau suivant indique le budget marketing de Google en millions de dollars en 2008 et 2011 (*The Wall Street Journal*, 27 mars 2012).

	2008	2011
Internet	26,0	123,3
Presse écrite	4,0	20,7
Télévision	0,0	69,3

a) Construire un diagramme en barres côte-à-côte avec l'année comme variable figurant sur l'axe horizontal. Commenter les tendances qui apparaissent.

b) Convertir le tableau ci-dessus en pourcentage alloué pour chaque année à chaque média. Construire un diagramme en barres empilées avec l'année comme variable figurant sur l'axe horizontal.

c) Quel graphique est le plus parlant ? Expliquer.

58. Un zoo a classé ses visiteurs en trois catégories : membre, école, et général. La catégorie « membre » fait référence aux visiteurs qui ont payé une redevance annuelle pour soutenir le zoo. Les membres bénéficient de certains avantages comme des remises sur les produits et les voyages organisés par le zoo. La catégorie « école » inclut les étudiants et les élèves des écoles primaires et secondaires. Ces visiteurs bénéficient généralement de tarifs réduits. La catégorie « général » inclut tous les autres visiteurs. Le zoo a récemment subi une baisse de fréquentation. Pour aider à mieux comprendre la fréquentation et l'adhésion des membres, un employé du zoo a collecté les données suivantes :

	Fréquentation			
Catégorie de visiteurs	2008	2009	2010	2011
Général	153 713	158 704	163 433	169 106
Membre	115 523	104 795	98 437	81 217
École	82 885	79 876	81 970	81 290
Total	352 121	343 375	343 840	331 613

a) Construire un diagramme en barres pour la fréquentation totale au cours du temps. Commenter toute tendance apparaissant dans les données.

b) Construire un diagramme en barres côte-à-côte montrant la fréquentation par catégorie de visiteurs avec l'année comme variable figurant sur l'axe horizontal.

c) Commenter l'évolution de la fréquentation du zoo en vous basant sur les graphiques construits aux questions (a) et (b).

PROBLÈME 1 *Les magasins Pelican*

Les magasins Pelican, une marque de National Clothing, sont une chaîne de magasins de vêtements pour femmes implantée à travers les États-Unis. Le magasin a récemment lancé

Tableau 2.19 ***Données d'un échantillon de 100 transactions réalisées dans les magasins Pelican***

Client	Type de client	Nombre d'articles	Montant d'achat	Moyen de paiement	Sexe	Statut marital	Âge
1	Régulier	1	39,50	Discover	Homme	Marié	32
2	Occasionnel	1	102,40	Carte de fidélité	Femme	Marié	36
3	Régulier	1	22,50	Carte de fidélité	Femme	Marié	32
4	Occasionnel	5	100,40	Carte de fidélité	Femme	Marié	28
5	Régulier	2	54,00	MasterCard	Femme	Marié	34
…	…	…	…	…	…	…	…
96	Régulier	1	39,50	MasterCard	Femme	Marié	44
97	Occasionnel	9	253,00	Carte de fidélité	Femme	Marié	30
98	Occasionnel	10	287,59	Carte de fidélité	Femme	Marié	52
99	Occasionnel	2	47,60	Carte de fidélité	Femme	Marié	30
100	Occasionnel	1	28,44	Carte de fidélité	Femme	Marié	44

une campagne de promotion en envoyant des bons de réduction aux clients des autres magasins National Clothing. Le fichier en ligne intitulé Magasins Pelican contient les données d'un échantillon de 100 transactions enregistrées au cours d'une journée dans les magasins Pelican alors que la campagne promotionnelle était en cours. Le tableau 2.19 reprend une partie du fichier. La méthode de paiement par carte de fidélité fait référence à des dépenses réglées en utilisant une carte National Clothing. Les clients qui font un achat en utilisant un bon de réduction sont référencés comme des clients occasionnels et les clients qui ont fait un achat mais n'ont pas utilisé un bon de réduction sont référencés comme des clients réguliers. Dans la mesure où les bons de réduction n'ont pas été envoyés aux clients réguliers des magasins Pelican, les responsables considèrent que les achats faits par les clients occasionnels n'auraient pas été réalisés en l'absence de bons de réduction. Bien sûr, les magasins Pelican espèrent que les clients occasionnels continueront à faire leurs achats dans leurs magasins. La plupart des variables présentées dans le tableau 2.17 sont explicites, mais deux variables nécessitent davantage d'explication.

Nombre d'articles : Nombre total d'articles achetés

Montant d'achat : Le montant total (en dollars) dépensés par carte de crédit

Les responsables des magasins Pelican souhaitent utiliser les données de cet échantillon pour mieux connaître leur base de clients et évaluer les politiques promotionnelles par bons de réduction.

Rapport

Utiliser les méthodes graphiques et sous forme de tableaux de statistiques descriptives pour définir le profil type des clients et évaluer l'impact de la campagne de promotion. Au minimum, votre rapport doit contenir :

1. Les distributions de fréquence en pourcentage des variables clés.
2. Un diagramme en barres ou un diagramme circulaire illustrant le pourcentage des achats attribuables à chaque moyen de paiement.
3. Une tabulation croisée du type de client (régulier ou occasionnel) et des achats. Commenter toutes similitudes ou différences observées.
4. Un nuage de points pour illustrer la relation entre les achats et l'âge des clients.

PROBLÈME 2 *L'industrie cinématographique*

L'industrie cinématographique est un secteur concurrentiel. Plus de 50 studios produisent globalement 300 à 400 films par an, et le succès financier de chaque film varie considérablement. Les recettes (en millions de dollars) lors du premier week-end après la sortie du film en salle, les recettes globales (en millions de dollars), le nombre de cinémas projetant le film et le nombre de semaines sur les écrans sont les variables généralement utilisées pour évaluer le succès d'un film. Les données collectées pour un échantillon de 100 films produits en 2011 sont regroupées dans le fichier en ligne intitulé Films 2011 (Box Office Mojo, 17 mars 2012). Le tableau 2.20 reprend les données pour les 10 premiers films de ce fichier.

Rapport

Utiliser les méthodes graphiques et sous forme de tableaux de statistiques descriptives pour déterminer comment ces variables contribuent au succès d'un film. Inclure les éléments suivants dans votre rapport.

Tableau 2.20 ***Données de performance pour 10 films***

Film	Recettes première semaine	Recettes totales	Nombre de cinémas projetant le film	Nombre de semaines sur les écrans
Harry Potter and the Deathly Hallows 2e Partie	169,19	381,01	4 375	19
Transformers : Dark of the Moon	97,85	352,39	4 088	15
The Twilight Saga: Breaking Dawn 1ère partie	138,12	281,29	4 066	14
The Hangover 2e partie	85,95	254,46	3 675	16
Pirates of the Caribbean : On Stranger Tide	90,15	241,07	4 164	19
Fast Five	86,20	209,84	3 793	15
Mission : Impossible - Ghost Protocol	12,79	208,55	3 555	13
Cars 2	66,14	191,45	4 115	25
Sherlock Holmes : A game of shadows	39,64	186,59	3 703	13
Thor	65,72	181,03	3 963	16

1. Des résumés graphiques et sous forme de tableaux de chacune des quatre variables, accompagnés d'une discussion sur ce que nous apprend chaque résumé sur l'industrie cinématographique.
2. Un nuage de points pour explorer la relation entre les recettes globales et les recettes réalisées lors du premier week-end de sortie en salle. Discuter.
3. Un nuage de points pour explorer la relation entre les recettes globales et le nombre de cinémas diffusant le film. Discuter.
4. Un nuage de points pour explorer la relation entre les recettes globales et le nombre de semaines sur les écrans. Discuter.

ANNEXE 2.1 UTILISER MINITAB POUR CONSTRUIRE DES PRÉSENTATIONS GRAPHIQUES ET SOUS FORME DE TABLEAUX

Minitab offre de nombreuses possibilités pour résumer des données sous forme de graphiques et de tableaux. Dans cette annexe, nous décrirons les étapes nécessaires à l'utilisation de Minitab pour créer un diagramme de points, un histogramme, un diagramme « stem-and-leaf » et un nuage de points.

A2.1.1 Diagramme de points

Nous utilisons les données sur la durée des audits, regroupées dans le tableau 2.4 (fichier en ligne Audit). Les données sur la durée des audits sont enregistrées dans la colonne C1 d'une feuille de calcul Minitab. Les étapes suivantes permettent de créer un diagramme de points.

Étape 1. Sélectionner le menu **Graph** et sélectionner **Dotplot**
Étape 2. Sélectionner **One Y, Simple** et cliquer sur **OK**
Étape 3. Quand la boîte de dialogue Dotplot-One Y, Simple apparaît :
Entrer C1 dans la boîte **Graph Variables**
Sélectionner **OK**

A2.1.2 Histogramme

Nous montrons les étapes de construction d'un histogramme, représentant les fréquences sur l'axe vertical, en utilisant les données sur la durée des audits du tableau 2.4 (fichier en ligne Audit). Les données figurent dans la colonne C1 d'une feuille de calcul Minitab. Pour obtenir un histogramme des données sur la durée des audits, les étapes suivantes sont nécessaires.

Étape 1. Sélectionner le menu **Graph**
Étape 2. Sélectionner **Histogram**

Étape 3. Quand la boîte de dialogue Histogram apparaît :
Sélectionner **Simple**
Étape 4. Quand la boîte de dialogue Histogram-Simple apparaît :
Entrer C1 dans la boîte **Graph variables**
Cliquer sur **OK**
Étape 5. Quand la boîte de dialogue Histogram apparaît :
Positionner la souris sur l'une des barres
Double-cliquer
Étape 6. Quand la boîte de dialogue Edit Bars apparaît :
Cliquer sur **Binning**
Sélectionner **Midpoint** sous **Interval Type**
Sélectionner **Midpoint/cutpoint positions** sous **Interval Definition**
Entrer **10:35/5** dans la boîte[3]
Cliquer sur **OK**

Notez que Minitab permet également de dimensionner l'axe des abscisses de façon à faire apparaître les valeurs numériques au centre des rectangles de l'histogramme. Si vous souhaitez obtenir cette fonction, modifiez l'étape 6 en y incluant la commande suivante : Sélectionner **Midpoint** pour le type d'intervalle et entrer 12:32/5 dans la boîte **Midpoint/ Cutpoint positions**. Ces étapes fournissent le même histogramme avec les centres des rectangles de l'histogramme nommés 12, 17, 22, 27 et 32.

A2.1.3 Diagramme « stem-and-leaf »

Nous utilisons les données relatives au test d'aptitude du tableau 2.8 pour illustrer la construction d'un diagramme « stem-and-leaf » (fichier en ligne Test d'aptitude). Les données figurent dans la colonne C1 d'une feuille de calcul Minitab. Les étapes suivantes génèrent le diagramme représenté dans la section 2.3.

Étape 1. Sélectionner le menu **Graph**
Étape 2. Sélectionner **Stem-and-leaf**
Étape 3. Quand la boîte de dialogue Stem-and-leaf apparaît :
Entrer C1 dans la boîte **Graph Variables**
Cliquer sur **OK**

A2.1.4 Nuage de points

Nous utilisons les données relatives au magasin d'équipement hi-fi du tableau 2.14 pour illustrer la construction d'un nuage de points (fichier en ligne Hi-fi). Les semaines sont numérotées de 1 à 10 dans la colonne C1, le nombre de spots publicitaires figure dans

3 Les étapes 5 et 6 sont optionnelles mais sont mentionnées ici pour montrer à l'utilisateur les possibilités offertes par Minitab pour construire l'histogramme. L'entrée 10:35/5 dans l'étape 6 indique que 10 est la valeur de départ pour la construction de l'histogramme, 35 est la valeur finale de l'histogramme et 5 correspond à la largeur de la classe.

la colonne C2, et les données sur les ventes dans la colonne C3 d'une feuille de calcul Minitab. Les étapes suivantes génèrent le nuage de points de la figure 2.7.

Étape 1. Sélectionner le menu **Graph**
Étape 2. Sélectionner **Scatterplot**
Étape 3. Sélectionner **Simple** et cliquer sur **OK**
Étape 4. Lorsque la boîte de dialogue Scatterplot-Simple apparaît :
Entrer C3 sous **Y variables** et C2 sous **X variables**
Cliquer sur **OK**

A2.1.5 Tabulation croisée

Nous utilisons les données sur les restaurants de Zagat, dont une partie figure dans le tableau 2.9 (fichier en ligne Restaurant). Les restaurants sont numérotés de 1 à 300 dans la colonne C1 d'une feuille de calcul Minitab. La colonne C2 contient les données relatives au niveau de qualité (bon, très bon et excellent) et la colonne C3 le prix du repas.

Minitab ne peut créer une tabulation croisée que pour des variables qualitatives. Or, le prix des repas est une variable quantitative. Il nous faut donc coder les données relatives aux prix des repas en spécifiant à quelle catégorie ils appartiennent. Les étapes suivantes permettent de coder les données sur les prix en créant quatre catégories de prix dans la colonne C4 : 10-19$, 20-29$, 30-39$ et 40-49$.

Étape 1. Sélectionner le menu **Data**
Étape 2. Sélectionner **Code**
Étape 3. Sélectionner **Numeric to Text**
Étape 4. Quand la boîte de dialogue Code – Numeric to Text apparaît :
Entrer C3 dans la boîte **Code data from columns**
Entrer C4 dans la boîte **Store coded data in columns**
Entrer 10:19 dans la première boîte **Original values** et 10-19$ dans la boîte adjacente **New**
Entrer 20:29 dans la seconde boîte **Original values** et 20-29$ dans la boîte adjacente **New**
Entrer 30:39 dans la troisième boîte **Original values** et 30-39$ dans la boîte adjacente **New**
Entrer 40:49 dans la quatrième boîte **Original values** et 40-49$ dans la boîte adjacente **New**
Cliquer sur **OK**

Pour chaque prix de la colonne C3, apparaît dans la colonne C4 la catégorie à laquelle ce prix est associé. On peut maintenant effectuer la tabulation croisée pour le niveau de qualité et le prix du repas en utilisant les données des colonnes C2 et C4. Les étapes suivantes permettent de créer une tabulation croisée similaire à celle fournie dans le tableau 2.10.

Étape 1. Sélectionner le menu **Stat**
Étape 2. Sélectionner **Tables**
Étape 3. Sélectionner **Cross Tabulation** et **Chi-Square**

Étape 4. Quand la boîte de dialogue apparaît :
Entrer C2 dans la boîte **For rows** et C4 dans la boîte **For columns**
Sélectionner **Counts** sous **Display**
Cliquer sur **OK**

ANNEXE 2.2 UTILISER EXCEL POUR CONSTRUIRE DES PRÉSENTATIONS GRAPHIQUES ET SOUS FORME DE TABLEAUX

Excel offre de nombreuses possibilités pour résumer des données sous forme de graphiques et de tableaux. Dans cette annexe, nous montrons comment utiliser Excel pour construire une distribution de fréquence, un diagramme en barres, un diagramme circulaire, un histogramme, un nuage de points et une tabulation croisée. Nous utiliserons trois des outils les plus performants d'Excel en matière d'analyse des données : la création de graphiques et la création de rapports à partir des fonctions Pivot Chart et PivotTable.

A2.2.1 Utiliser Excel pour construire une distribution de fréquence, une distribution de fréquence relative et une distribution de fréquence en pourcentage

Nous pouvons utiliser l'outil Excel « PivotTables » pour construire une distribution de fréquence de l'échantillon des 50 achats de boisson non alcoolisée. Ouvrez le fichier en ligne intitulé Boisson non alcoolisée. Les données sont contenues dans les cellules A2:A51 et sont nommées dans la cellule A1.

Les étapes suivantes décrivent comment utiliser l'outil Excel « PivotTables » pour construire une distribution de fréquence de l'échantillon des 50 achats de boisson non alcoolisée.

Étape 1. Sélectionner une cellule de l'ensemble de données
Étape 2. Cliquer sur **Insert** dans la barre des tâches
Étape 3. Dans **Tables Group** choisir **Recommended PivotTables** ; une prévisualisation montrant la distribution de fréquence apparaît
Étape 4. Cliquer sur **OK** ; la distribution de fréquence apparaît dans une nouvelle feuille de calcul

La feuille de calcul de la figure 2.16 montre la distribution de fréquence pour les 50 achats de boisson non alcoolisée créée en suivant ces étapes. La boîte de dialogue PivotTable Fields, un élément clé des rapports PivotTable, est également présentée. Nous discuterons plus tard de l'utilisation de la boîte de dialogue PivotTable Fields dans l'annexe.

Options d'édition Vous pouvez facilement modifier le titre des colonnes dans l'output de la distribution de fréquence. Par exemple, pour changer le titre actuel

qui apparaît dans la cellule A3 (Titre des lignes) en « Boisson non alcoolisée », cliquer sur la cellule A3 et taper « Boisson non alcoolisée » ; pour modifier le titre de la cellule B3 (Somme des marques achetées) en « Fréquence », cliquez sur la cellule B3 et taper

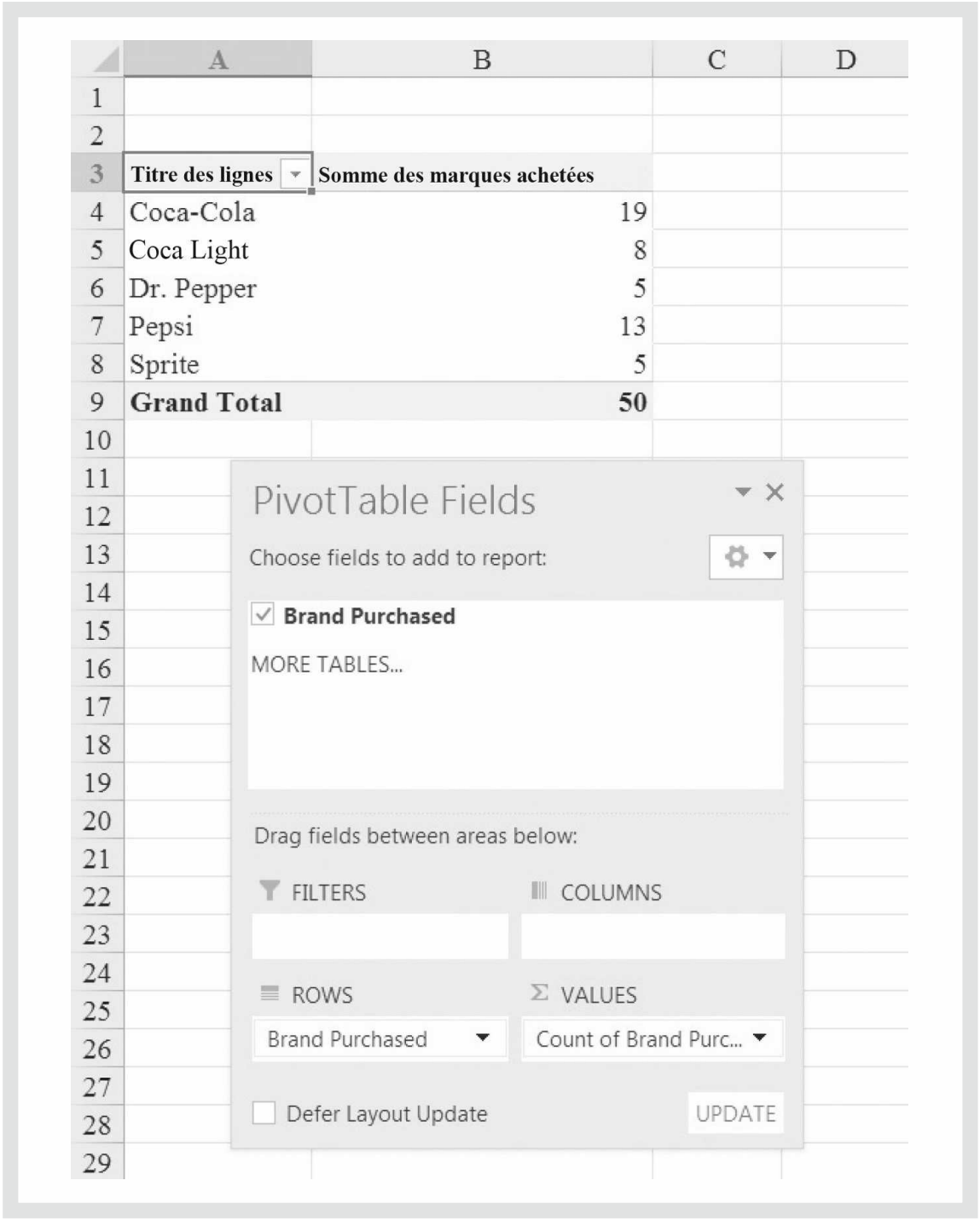


Figure 2.16 ***Distribution de fréquence pour les achats de boisson non alcoolisée construite en utilisant l'outil « Recommended PivotTables » d'Excel***

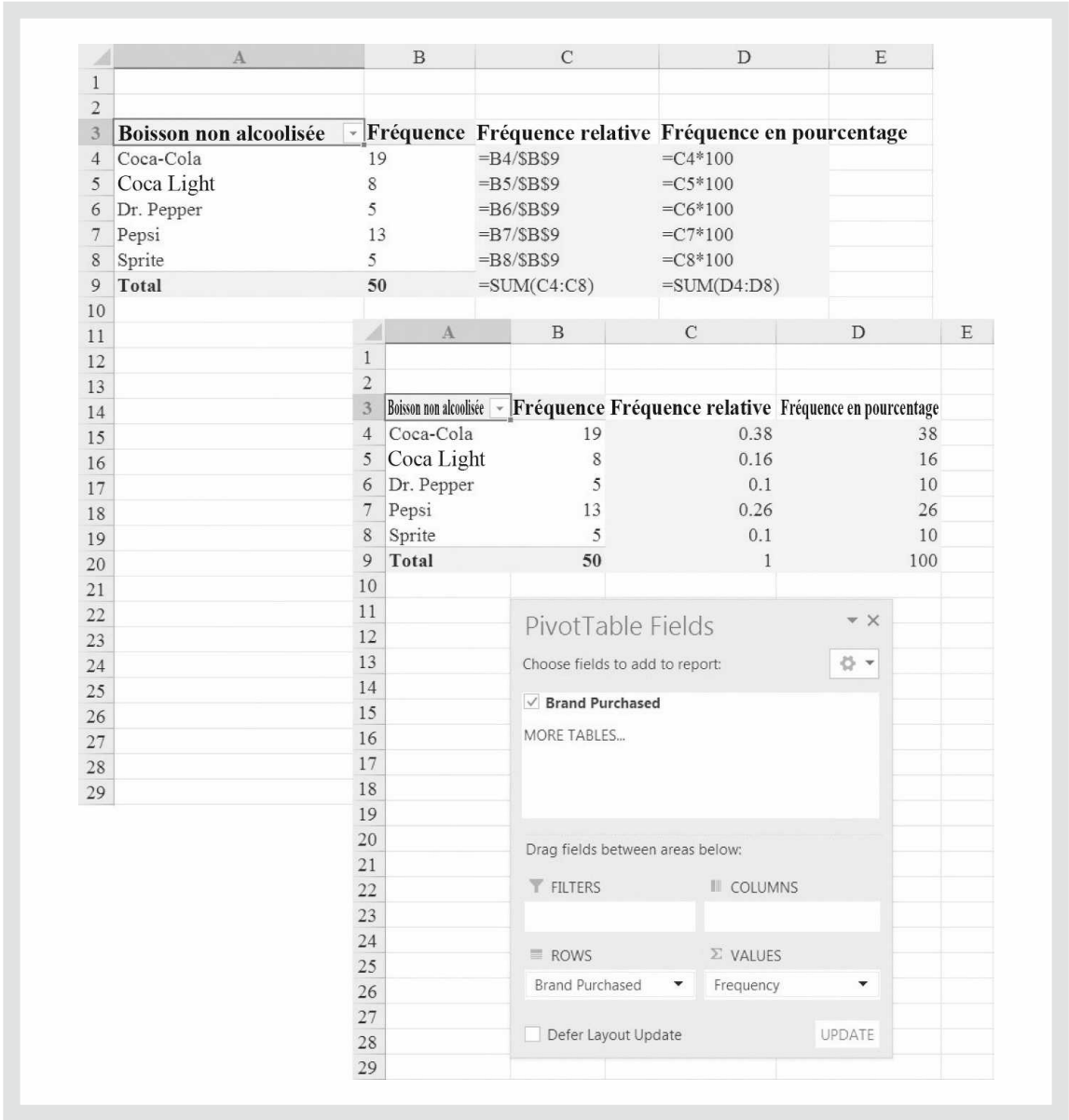


Figure 2.17 ***Distributions de fréquence relative et en pourcentage pour les achats de boisson non alcoolisée construites en utilisant les fonctions d'Excel***

« Fréquence » ; et pour modifier le titre de la cellule A9 (Grand Total) en « Total », cliquer sur la cellule A9 et taper « Total ». Les feuilles de calcul apparaissant au premier plan et en arrière-plan à la figure 2.17 contiennent les titres révisés ; en plus, le titre « Fréquence relative » a été entré dans la cellule C3 et le titre « Fréquence en pourcentage » a été ajouté dans la cellule D3 pour illustrer comment calculer les distributions de fréquence relative et en pourcentage.

Entrer des fonctions et des formules Référez-vous à la figure 2.17 pour suivre nos indications pour créer des distributions de fréquence relative et en pourcentage

pour les achats de boisson non alcoolisée. La feuille de calcul contenant les formules se trouve en arrière-plan et la feuille fournissant les résultats au premier plan. Pour calculer la fréquence relative pour Coca-Cola en utilisant l'équation (2.1), nous avons entré la formule =B4/B9 dans la cellule C4 ; le résultat, 0,38, correspond à la fréquence relative pour Coca-Cola. Copier la cellule C4 dans les cellules C5:C8 permet de calculer les fréquences relatives pour chacune des autres boissons non alcoolisées. Pour calculer la fréquence en pourcentage pour Coca-Cola, nous avons entré la formule =C4*100 dans la cellule D4. Le résultat, 38, indique que 38 % des achats de boisson non alcoolisée se sont portés sur la marque Coca-Cola. Copier la cellule D4 dans les cellules D5:D8 permet de calculer les fréquences en pourcentage pour chacune des autres marques de boisson non alcoolisée. Pour calculer le total des fréquences relatives, nous avons entré la formule =SUM(C4:C8) dans la cellule C9. Et pour calculer le total des fréquences en pourcentage, nous avons copié la cellule C9 dans la cellule C10.

A2.2.2 Utiliser Excel pour construire un diagramme en barres et un diagramme circulaire

Nous pouvons utiliser l'outil Excel « Recommended Charts » pour construire un diagramme en barres et un diagramme circulaire pour l'échantillon des 50 achats de boisson non alcoolisée. Ouvrez le fichier en ligne intitulé Boisson non alcoolisée. Les données sont contenues dans les cellules A2:A51 et sont nommées dans la cellule A1.
Les étapes suivantes décrivent comment utiliser l'outil Excel « Recommended Charts » pour construire un diagramme en barres pour l'échantillon des 50 achats de boisson non alcoolisée.

Étape 1. Sélectionner une cellule de l'ensemble de données
Étape 2. Cliquer sur **Insert** dans la barre des tâches
Étape 3. Dans **Charts Group** choisir **Recommended Charts** ; une pré-visualisation montrant le graphique apparaît
Étape 4. Cliquer sur **OK** ; le diagramme en barres apparaît dans une nouvelle feuille de calcul

La feuille de calcul de la figure 2.18 montre le diagramme en barres pour les 50 achats de boisson non alcoolisée, créé en suivant ces étapes. La fréquence de distribution et la boîte de dialogue PivotTable Fields, créées par Excel pour construire le diagramme en barres, apparaissent également. Ainsi, en utilisant l'outil « Recommended Charts » d'Excel, vous pouvez construire un diagramme en barres et une distribution de fréquence en même temps.

> Le diagramme en barres de la figure 2.18 est référencé par Excel sous le terme « Clustered Column chart ».

Options d'édition Vous pouvez facilement modifier le titre du diagramme en barres et nommer les axes. Par exemple, supposez que vous vouliez nommer le graphique de la façon suivante : « Diagramme en barres des achats de boisson

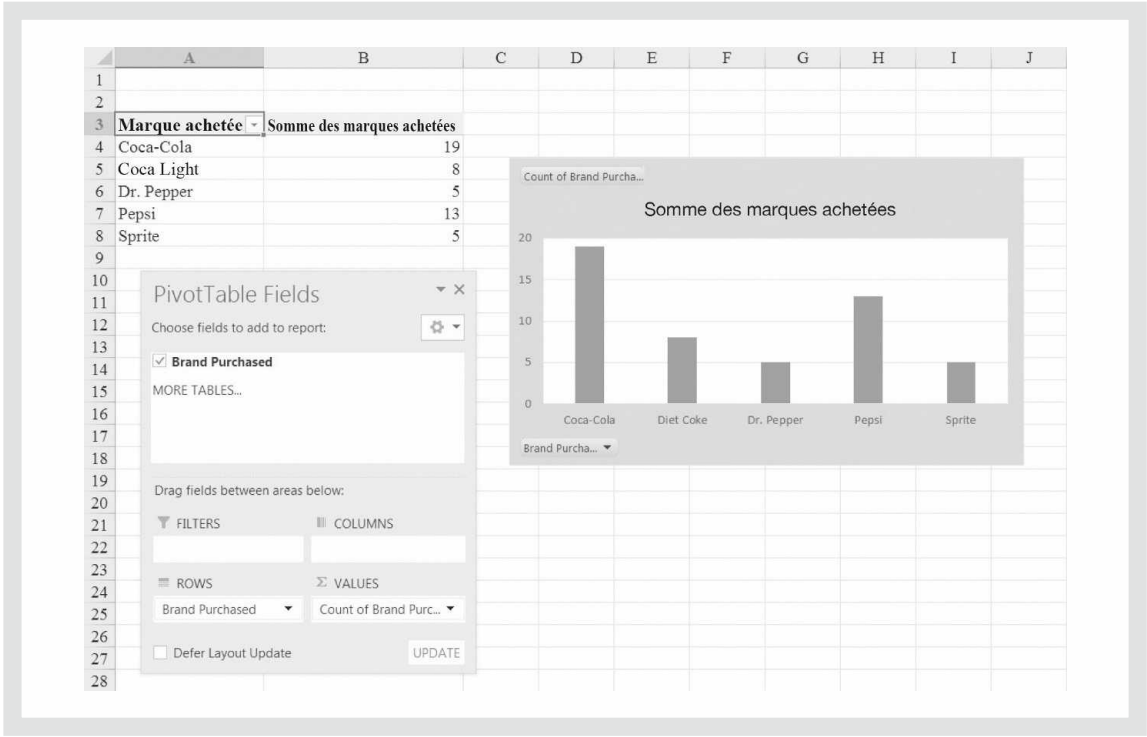


Figure 2.18 ***Diagramme en barres des achats de boisson non alcoolisée construit en utilisant l'outil « Recommended Charts » d'Excel***

non alcoolisée » et insérer les titres « Boisson non alcoolisée » sur l'axe horizontal et « Fréquence » sur l'axe vertical.

Étape 1. Cliquer sur **Chart Title** et remplacer-le par **Diagramme en barres des achats de boisson non alcoolisée**

Étape 2. Cliquer sur le bouton **Chart Elements** + (situé à côté du coin supérieur droit du graphique)

Étape 3. Lorsque la liste des éléments du graphique apparaît :
Cliquer sur **Axis Title** (crée un espace pour inscrire un titre sur les axes)

Étape 4. Cliquer sur **Horizontal (Category) Axis Title** et remplacer-le par **Boisson non alcoolisée**

Étape 5. Cliquer sur **Vertical (Value) Axis Title** et remplacer-le par **Fréquence**

Le diagramme en barres modifié apparaît à la figure 2.19.

Créer un diagramme circulaire Pour créer un diagramme circulaire, sélectionner le diagramme en barres (en cliquant n'importe où sur le graphique) pour faire apparaître trois tableaux (**Analyze**, **Design** et **Format**) situé sur la barre des tâches sous le titre **PivotChart Tools**. Cliquer sur **Design Tab** et choisir l'option **Change Chart Type** pour faire apparaître la boîte de dialogue. Cliquer sur l'option **Pie** et ensuite sur **OK** pour faire apparaître le diagramme circulaire des achats de boisson non alcoolisée.

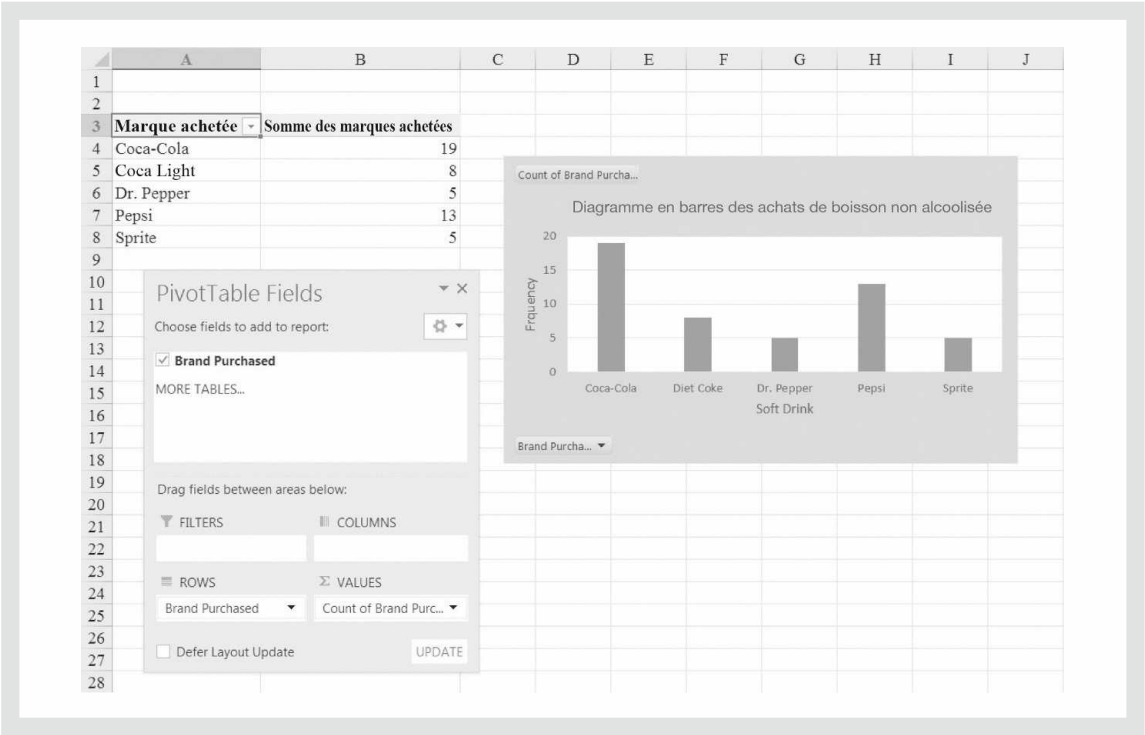


Figure 2.19 ***Diagramme en barres modifié des achats de boisson non alcoolisée construit en utilisant l'outil « Recommended Charts » d'Excel***

A2.2.3 Utiliser Excel pour construire une distribution de fréquence

Précédemment, nous avons illustré comment utiliser l'outil « Recommended PivotTables » d'Excel pour construire une distribution de fréquence. Nous pouvons également utiliser directement l'outil PivotTable d'Excel pour cela. Nous illustrons la marche à suivre avec les données sur la durée des audits. Ouvrez le fichier en ligne intitulé Audit. Les données apparaissent dans les cellules A2:A21 et un nom dans la cellule A1.

Les étapes suivantes décrivent comment utiliser l'outil PivotTable d'Excel pour construire une distribution de fréquence à partir des données sur la durée des audits. Lorsqu'on utilise l'outil PivotTable d'Excel, chaque colonne de données correspond à un champ. Ainsi, dans l'exemple sur la durée des audits, les données apparaissant dans les cellules A2:A21 et le nom figurant dans la cellule A1 sont référencés sous le terme « champ des durées d'audit ».

Étape 1. Sélectionner une cellule dans l'ensemble de données (cellules A1:A21)
Étape 2. Cliquer sur **Insert** dans la barre des tâches
Étape 3. Dans le groupe **Tables**, choisir **PivotTable**
Étape 4. Lorsque la boîte de dialogue **Create PivotTable** apparaît :

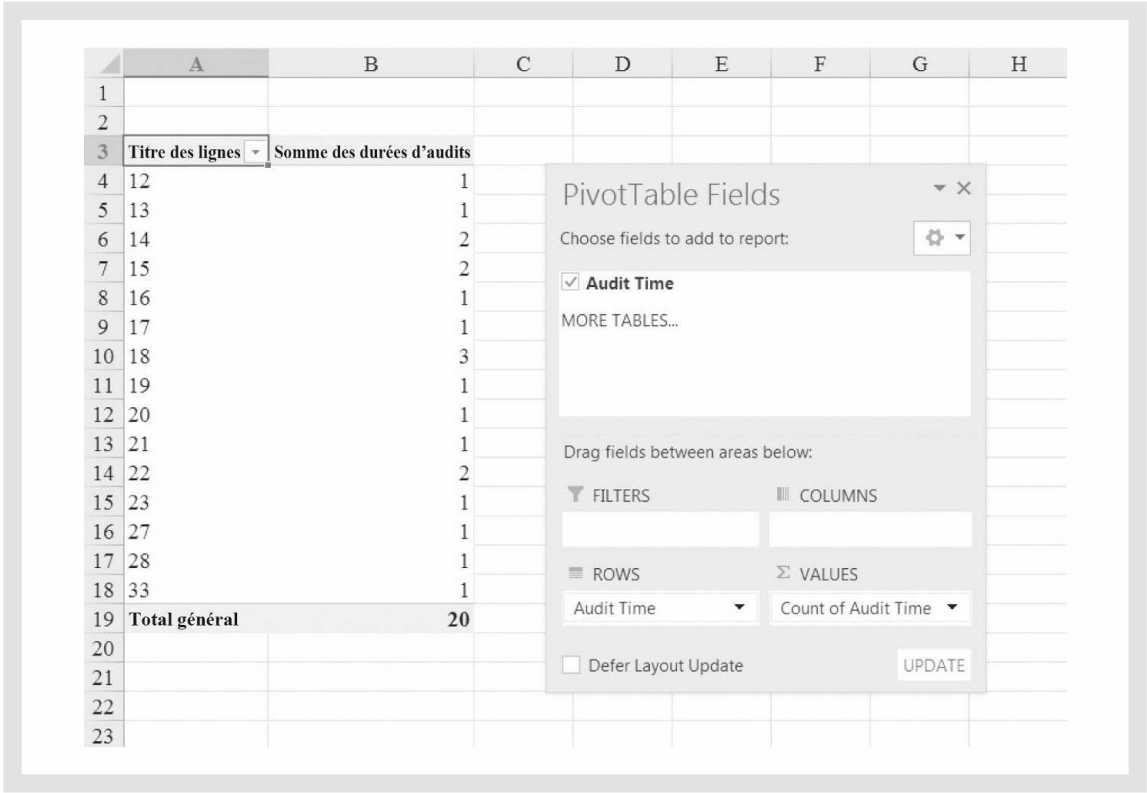


Figure 2.20 ***Liste PivotTable Fields et la PivotTable initiale utilisée pour construire une distribution de fréquence pour les données sur la durée des audits***

Cliquer sur **OK** ; une boîte de dialogue apparaît dans une nouvelle feuille de calcul

Étape 5. Dans la boîte de dialogue **PivotTable Field** :
Déplacer le champ **Audit Time** vers la zone **Rows**
Déplacer le champ **Audit Time** vers la zone **Values**

Étape 6. Cliquer sur **Sum of Audit Time** dans la zone **Values**

Étape 7. Cliquer sur **Value Field Settings** dans la liste d'options qui apparaît

Étape 8. Lorsque la boîte de dialogue Value Field Settings
Sous **Summarize value field by**, choisir **Count**
Cliquer sur **OK**

La figure 2.20 représente la liste PivotTable Fields qui en résulte et la PivotTable correspondante. Pour construire la distribution de fréquence présentée dans le tableau 2.5, nous devons regrouper les lignes contenant les durées d'audits. Les étapes suivantes permettent de le faire.

Étape 1. Cliquer-droit sur la cellule A4 dans la PivotTable ou sur une autre cellule contenant une durée d'audit

Étape 2. Choisir **Group** dans la liste d'options qui apparaît

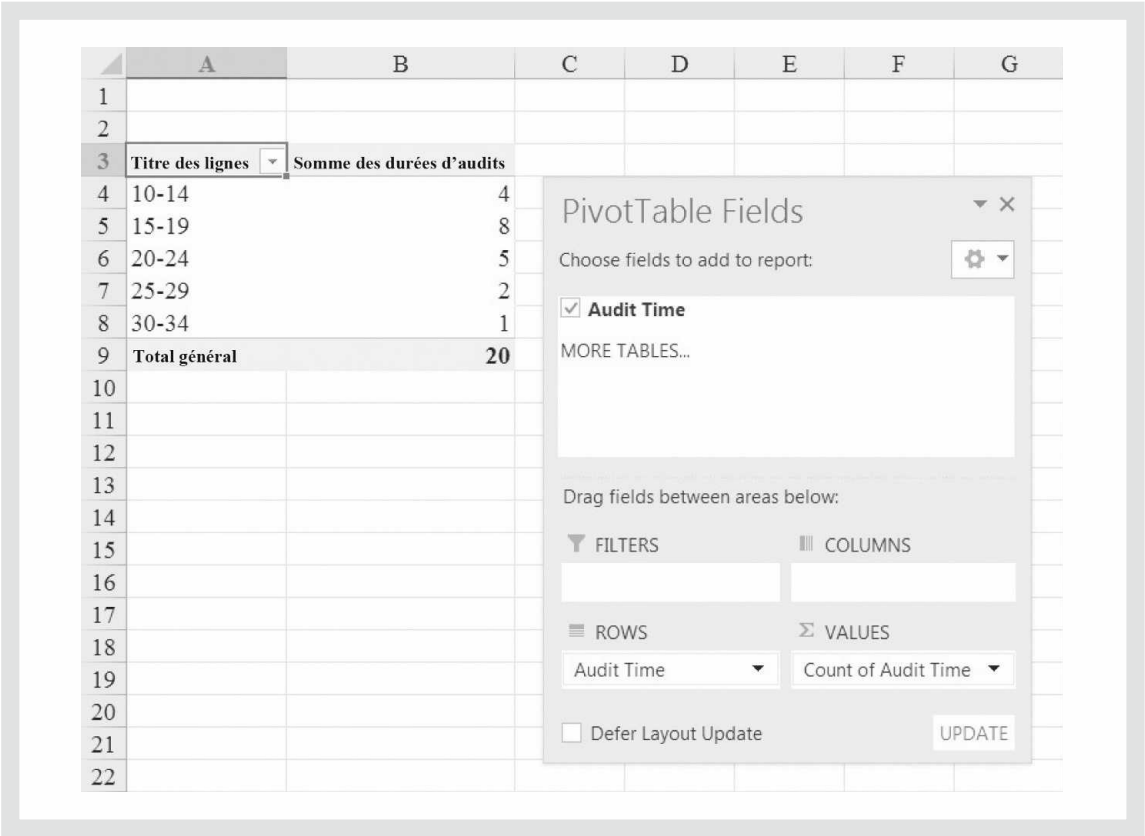


Figure 2.21 ***Distribution de fréquence pour les données sur la durée des audits construite en utilisant l'outil PivotTable d'Excel***

Étape 3. Lorsque la boîte de dialogue Grouping apparaît :
Entrer 10 dans la boîte **Starting at**
Entrer 34 dans la boîte **Ending at**
Entrer 5 dans la boîte **By**
Cliquer sur **OK**

La figure 2.21 présente la liste complète de PivotTable Fields et la PivotTable correspondante. Nous voyons qu'à l'exception des titres des colonnes, la PivotTable fournit les mêmes informations que la distribution de fréquence présentée dans le tableau 2.5.

Options d'édition Vous pouvez facilement modifier les noms figurant dans la PivotTable et les remplacer par ceux figurant dans le tableau 2.5. Par exemple, pour changer l'intitulé de la cellule A3 (Titre des lignes) par « Durée des audits (en jours) », cliquer sur la cellule A3 et taper « Durée des audits (en jours) » ; pour changer l'intitulé de la cellule B3 (Somme des durées d'audits) en « Fréquence », cliquer sur la cellule B3 et taper « Fréquence » ; et pour changer l'intitulé de la cellule A9 (Total général) en « Total », cliquer sur la cellule A9 et taper « Total ».

Les mêmes procédures suivies dans la première section de cette annexe peuvent maintenant être appliquées pour développer les distributions de fréquence relative et en pourcentage.

A2.2.4 Utiliser l'outil « Recommended Charts » d'Excel pour construire un histogramme

Dans la figure 2.21, nous avons montré les résultats obtenus en utilisant l'outil PivotTable d'Excel pour construire une distribution de fréquence pour les données sur la durée des audits. Nous utiliserons ces résultats pour illustrer comment l'outil « Recommended Charts » d'Excel peut être utilisé pour construire un histogramme décrivant les données quantitatives résumées dans une distribution de fréquence. Référez-vous à la figure 2.21 pour suivre les étapes.

Les étapes suivantes décrivent comment utiliser l'outil « Recommended Charts » d'Excel pour construire un histogramme pour les données sur la durée des audits.

Étape 1. Sélectionner une cellule dans le rapport PivotTable (cellules A3:B9 de la figure 2.21)

Étape 2. Cliquer sur **Insert** dans la barre des tâches

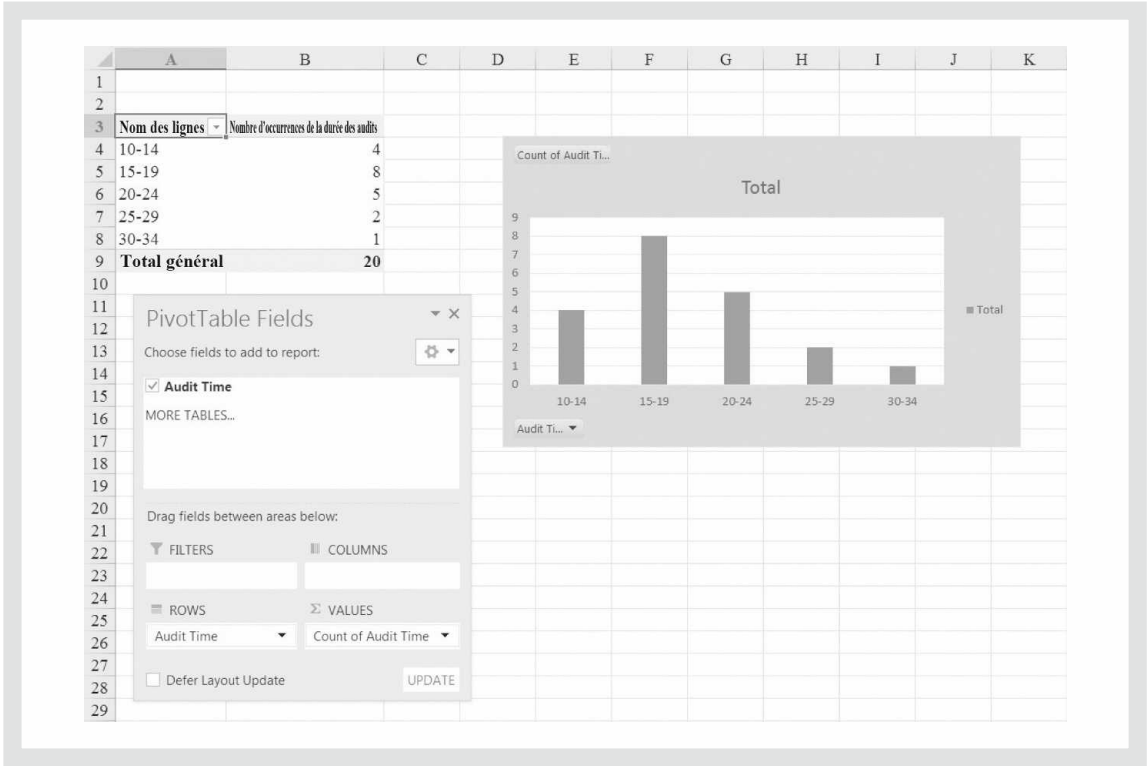


Figure 2.22 *Graphique initial utilisé pour construire un histogramme des données sur la durée des audits*

Étape 3. Dans le groupe **Charts**, choisir **Recommended Charts** ; une prévisualisation du graphique apparaît

Étape 4. Cliquer **OK**

La feuille de calcul de la figure 2.22 représente le graphique pour les données sur la durée des audits créé en suivant ces étapes. À l'exception des espaces séparant les barres, il ressemble à l'histogramme pour les données sur la durée des audits présenté à la figure 2.5. Nous pouvons facilement modifier ce graphique pour supprimer les espaces entre les barres et entrer des intitulés pour les axes et un titre plus pertinents.

Options d'édition En plus de supprimer les espaces entre les barres, supposez que vous souhaitez modifier le titre du graphique et le nommer « Histogramme des données sur la durée des audits » et insérer l'intitulé « Durée des audits (en jours) » sur l'axe horizontal et « Fréquence » sur l'axe vertical.

Étape 1. Cliquer-droit sur une barre du graphique et choisir **Format Data Series** dans la liste d'options qui apparaît

Étape 2. Lorsque la boîte de dialogue apparaît :

Aller à la section **Series Options**

Fixer **Gap Width** à 0

Cliquer sur le bouton **Close** en haut à droite de la boîte de dialogue

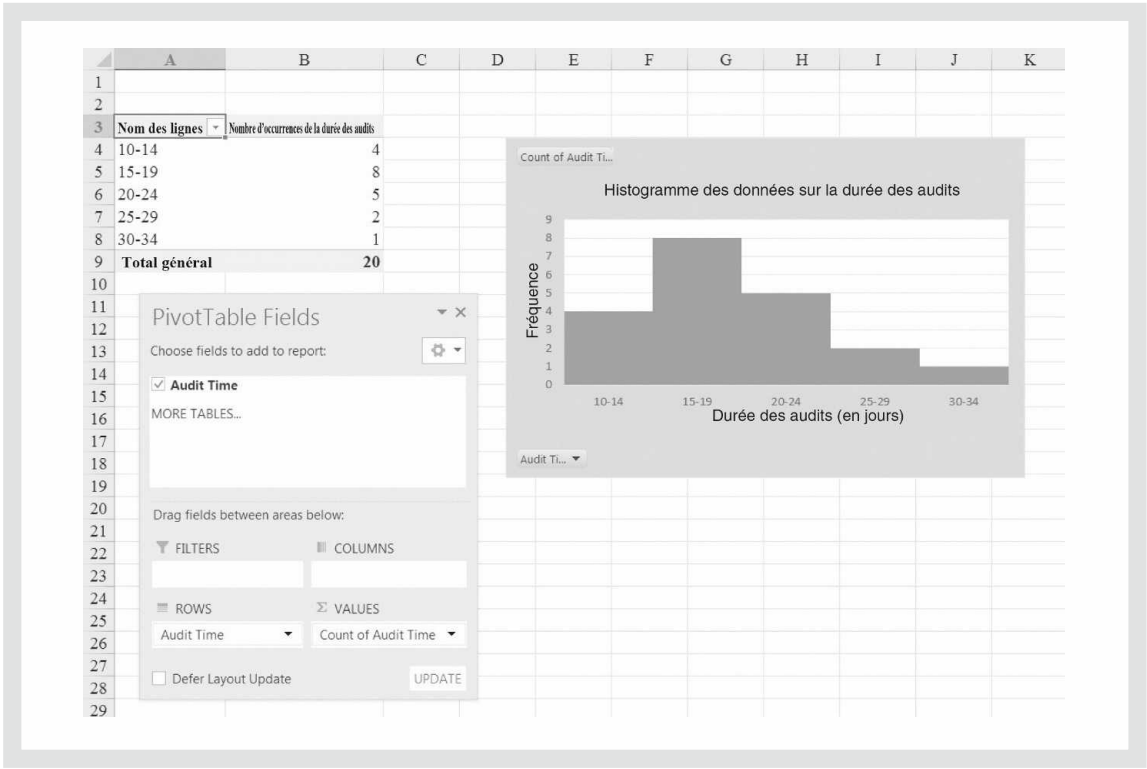


Figure 2.23 *Histogramme des données sur la durée des audits, créé en utilisant l'outil « Recommended Charts » d'Excel*

Étape 3. Cliquer sur **Chart Title** et le remplacer par **Histogramme des données sur la durée des audits**

Étape 4. Cliquer sur le bouton **Chart Elements** (situé à côté du coin supérieur droit du graphique)

Étape 5. Lorsque la liste des éléments du graphique apparaît :
Cliquer sur **Axis Titles** (crée un espace pour les titres des axes)
Cliquer sur **Legends** pour décocher l'élément dans la boîte Legends

Étape 6. Cliquer sur **Horizontal (Category) Axis Title** et le remplacer par **Durée des audits (en jours)**

Étape 7. Cliquer sur **Vertical (Value) Axis Title** et le remplacer par **Fréquence**

L'histogramme modifié pour la durée des audits apparaît à la figure 2.23.

A2.2.5 Utiliser l'outil PivotTable d'Excel pour construire une tabulation croisée

Restaurant

L'outil PivotTable d'Excel peut être utilisé pour résumer les données relatives à au moins deux variables simultanément. Nous illustrerons l'utilisation de cet outil en montrant comment effectuer une tabulation croisée du rapport qualité/prix des repas à partir des données sur 300 restaurants de Los Angeles. Ouvrez le fichier en ligne Restaurant. Les données

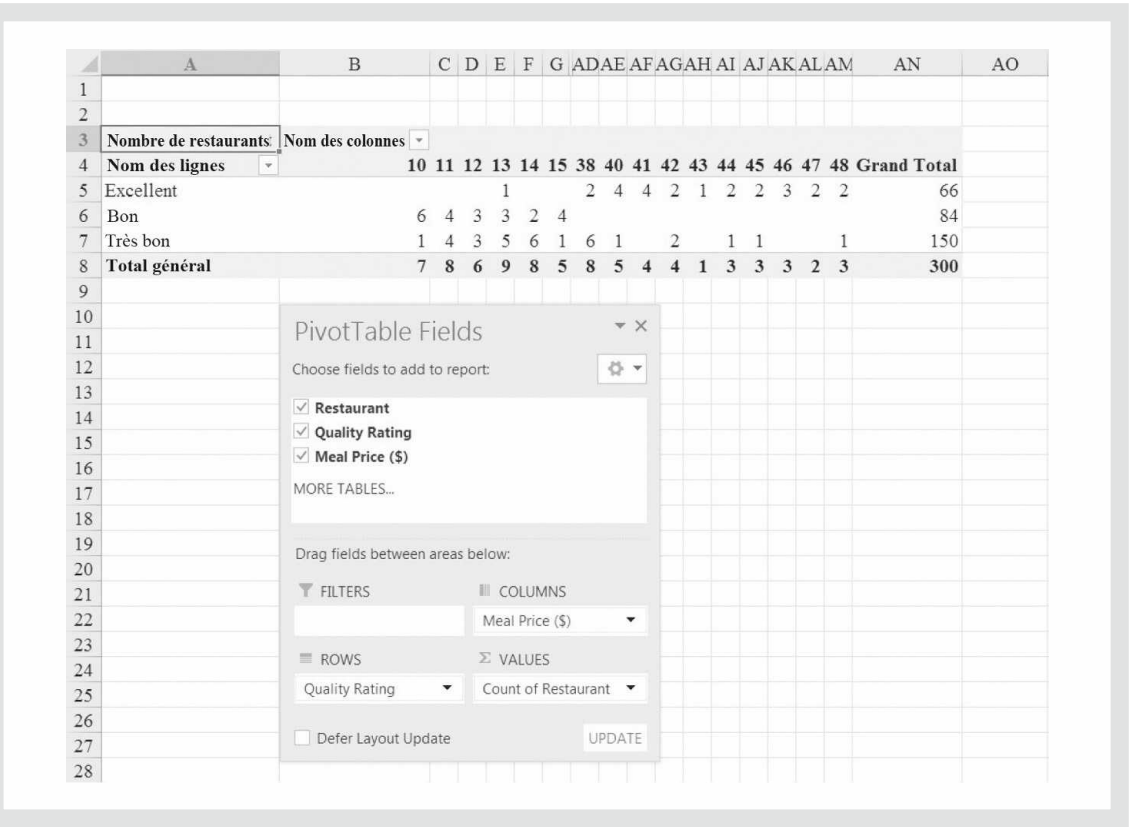


Figure 2.24 *Boîte de dialogue PivotTable Fields initiale et PivotTable pour les données sur les restaurants*

sont enregistrées dans les cellules B2:C301 et les intitulés figurent dans la colonne A et les cellules B1:C1.

Chacune des trois colonnes de l'ensemble de données Restaurant, intitulées « Restaurant », « Niveau de qualité » et « Prix du repas ($) » correspond à un champ. Les champs peuvent être choisis pour représenter des lignes, des colonnes ou des valeurs dans la PivotTable. Les étapes suivantes décrivent comment utiliser l'outil PivotTable d'Excel pour construire une tabulation croisée des niveaux de qualité et du prix des repas.

Étape 1. Sélectionner la cellule A1 ou toute autre cellule dans l'ensemble de données
Étape 2. Cliquer sur **Insert** dans la barre des tâches
Étape 3. Dans le groupe **Tables**, choisir **PivotTable**
Étape 4. Quand la boîte de dialogue Create PivotTable apparaît :

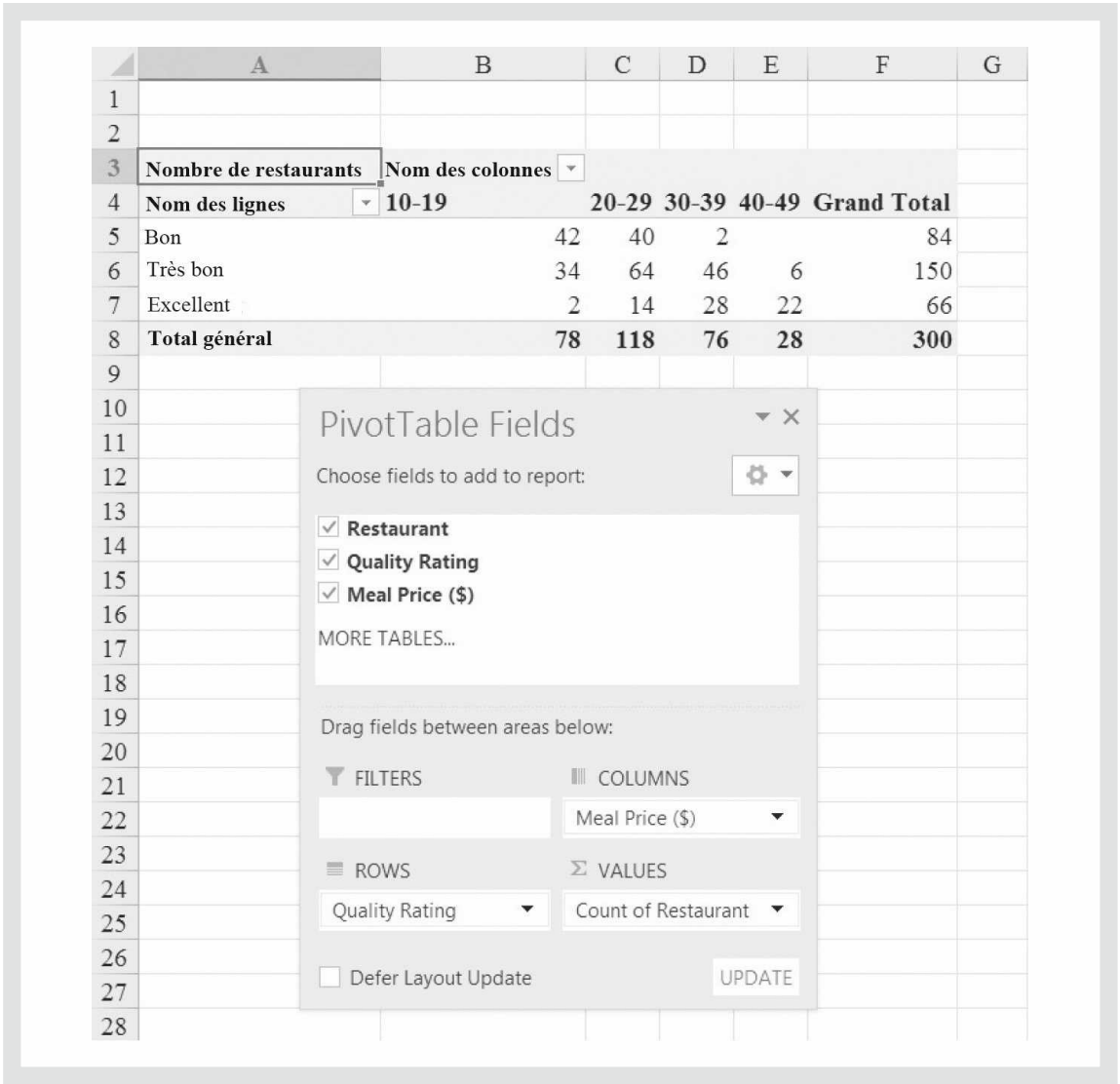


Figure 2.25 *PivotTable finale pour les données sur les restaurants*

Cliquer sur **OK** et une PivotTable ainsi que la boîte de dialogue apparaissent

Étape 5. Dans la boîte de dialogue PivotTable Fields :
Déplacer le champ **Niveau de qualité** vers la zone **Rows**
Déplacer le champ **Prix du repas** dans la zone **Columns**
Déplacer le champ **Restaurant** vers la zone **Values**

Étape 6. Cliquer sur **Sum of Restaurant** dans la zone **Values**

Étape 7. Cliquer sur **Value Field Settings** dans la liste d'options qui apparaît

Étape 8. Lorsque la boîte de dialogue apparaît :
Sous **Summarize value field by**, choisir **Count**
Cliquer sur **OK**

La figure 2.24 montre la liste PivotTable Fields et la PivotTable correspondante créée en suivant ces étapes. Pour des questions de lisibilité, les colonnes H:AC ont été masquées.

Options d'édition Pour compléter la PivotTable, nous devons regrouper les lignes contenant les prix des repas et ordonner correctement les niveaux de qualité. Les étapes suivantes permettent cela.

Étape 1. Cliquer-droit sur la cellule B4 dans la PivotTable ou sur toute autre cellule contenant les prix des repas

Étape 2. Choisir **Group** dans la liste d'options qui apparaît

Étape 3. Lorsque la boîte de dialogue apparaît :
Entrer 10 dans la boîte **Starting at**
Entrer 49 dans la boîte **Ending at**
Entrer 10 dans la boîte **By**
Cliquer sur **OK**

Étape 4. Cliquer-droit sur **Excellent** dans la cellule A5

Étape 5. Choisir **Move** et cliquer sur **Move « Excellent » to End**

La PivotTable finale apparaît dans la figure 2.25. Notez qu'elle fournit la même information que la tabulation croisée présentée dans le tableau 2.10.

A2.2.6 Utiliser l'outil Charts d'Excel pour créer un nuage de points et une droite de tendance

Nous pouvons utiliser l'outil Charts d'Excel pour créer un nuage de points et une droite de tendance pour les données relatives au magasin d'équipement hi-fi. Ouvrez le fichier en ligne intitulé Hi-fi. Les données sont enregistrées dans les cellules B2:C11 et les intitulés sont notés dans la colonne A et les cellules B1:C1.

Les étapes suivantes décrivent comment utiliser l'outil Charts d'Excel pour créer un nuage de points à partir des données contenues dans la feuille de calcul.

Étape 1. Sélectionner les cellules B1:C11

Étape 2. Cliquer sur **Insert** dans la barre des tâches

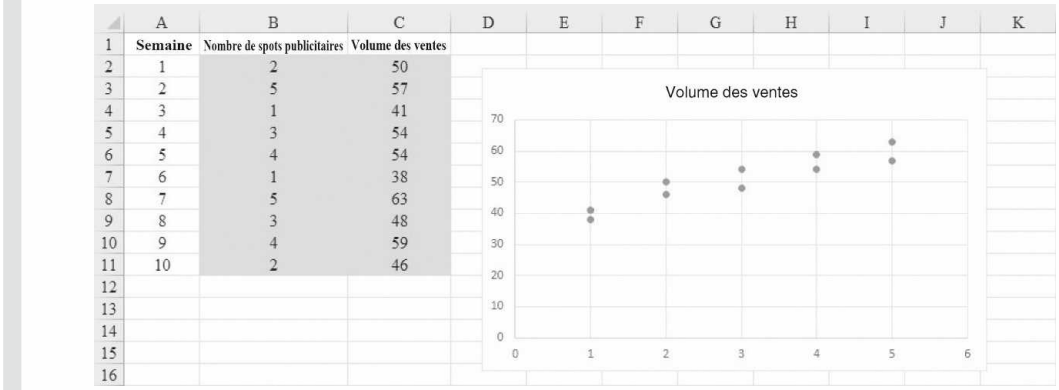


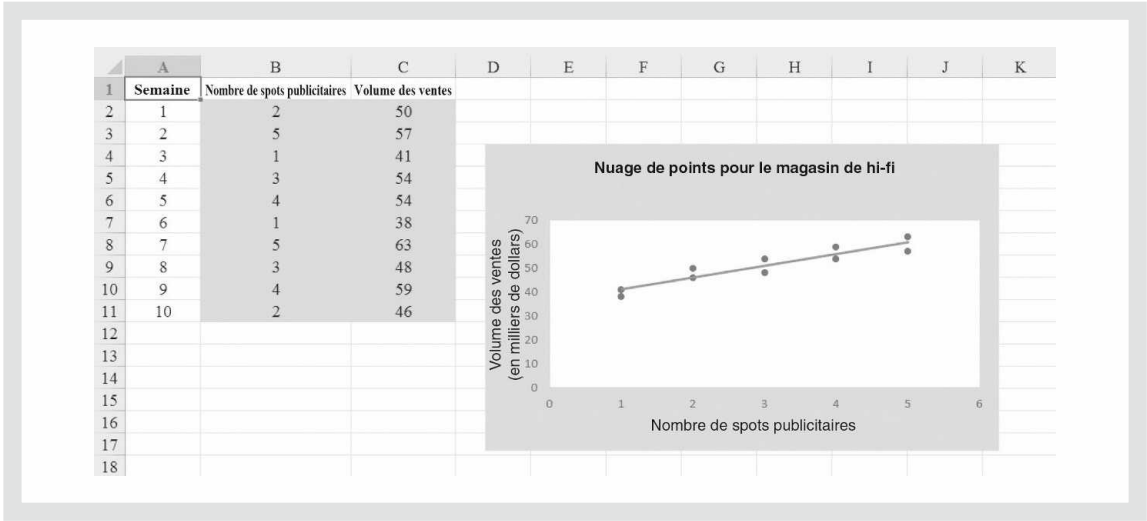
Figure 2.26 *Nuage de points initial pour les données relatives au magasin d'équipements hi-fi obtenu en utilisant l'outil Recommended Charts d'Excel*

Étape 3. Dans le groupe **Charts**, cliquer sur **Insert Scatter (X,Y)** ou **Bubble Chart**
Étape 4. Lorsque la liste des différents nuages de points apparaît :
Cliquer sur **Scatter** (le graphique dans le coin supérieur gauche)

La feuille de calcul de la figure 2.26 représente le nuage de points créé en suivant ces instructions.

Options d'édition Vous pouvez aisément modifier le nuage de points pour faire apparaître un titre de graphique différent, nommer les axes et faire apparaître une droite de tendance. Par exemple, supposez que vous vouliez nommer le graphique « Nuage de points pour le magasin de hi-fi », l'axe horizontal « Nombre de spots publicitaires » et l'axe vertical « Ventes (en milliers de dollars) ».

Étape 1. Cliquer sur **Chart Title** et remplacer-le par **Nuage de points pour le magasin de hi-fi**
Étape 2. Cliquer sur le bouton Chart Elements (situé à côté du coin supérieur droit du graphique)
Étape 3. Lorsque la liste des éléments apparaît :
Cliquer sur **Axis Title** (crée un endroit pour y faire figurer les titres des axes)
Cliquer sur **Gridlines** (pour déselectionner l'option Gridlines)
Cliquer sur **Trendline**
Étape 4. Cliquer sur **Horizontal (Value) Axis Title** et remplacer-le par **Nombre de spots publicitaires**
Étape 5. Sélectionner **Vertical (Value) Axis Title** et remplacer-le par **Volume des ventes (en milliers de dollars)**
Étape 6. Pour passer d'une droite de tendance en pointillé à une droite en trait plein, cliquer-droit sur la droite de tendance et sélectionner l'option **Format Trendline**



Semaine	Nombre de spots publicitaires	Volume des ventes
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

Figure 2.27 ***Nuage de points et droite de tendance modifiés pour le magasin de hi-fi créés en utilisant l'outil Recommended Charts d'Excel***

Étape 7. Lorsque la boîte de dialogue apparaît :
Sélectionner l'option **Fill & Line**
Dans la boîte **Dash type**, sélectionner **Solid**
Fermer la boîte de dialogue

Le nuage de points et la droite de tendance modifiés sont présentés à la figure 2.27.

A2.2.7 Utiliser l'outil Recommended Charts d'Excel pour construire des diagrammes en barres côte-à-côte et empilées

Restaurant

À la figure 2.25, nous avons montré les résultats obtenus en utilisant l'outil PivotTable d'Excel pour construire une distribution de fréquence pour l'échantillon des 300 restaurants situés autour de Los Angeles. Nous utilisons ces résultats pour illustrer comment utiliser l'outil Recommended Charts d'Excel pour construire des diagrammes en barres côte-à-côte et empilées pour les données sur les restaurants en utilisant l'output PivotTable.

Les étapes suivantes décrivent comment utiliser l'outil Recommended Charts d'Excel pour construire un diagramme en barres côte-à-côte pour les données sur les restaurants en utilisant l'output de l'outil PivotTable présenté à la figure 2.25.

Étape 1. Sélectionner une cellule dans le rapport PivotTable (cellules A3:F8 de la figure 2.25)
Étape 2. Cliquer sur **Insert** dans la barre des tâches
Étape 3. Dans le Groupe **Charts**, choisir **Recommended Charts** ; une prévisualisation d'un diagramme en barres avec les niveaux de qualité sur l'axe horizontal apparaît
Étape 4. Cliquer sur **OK**

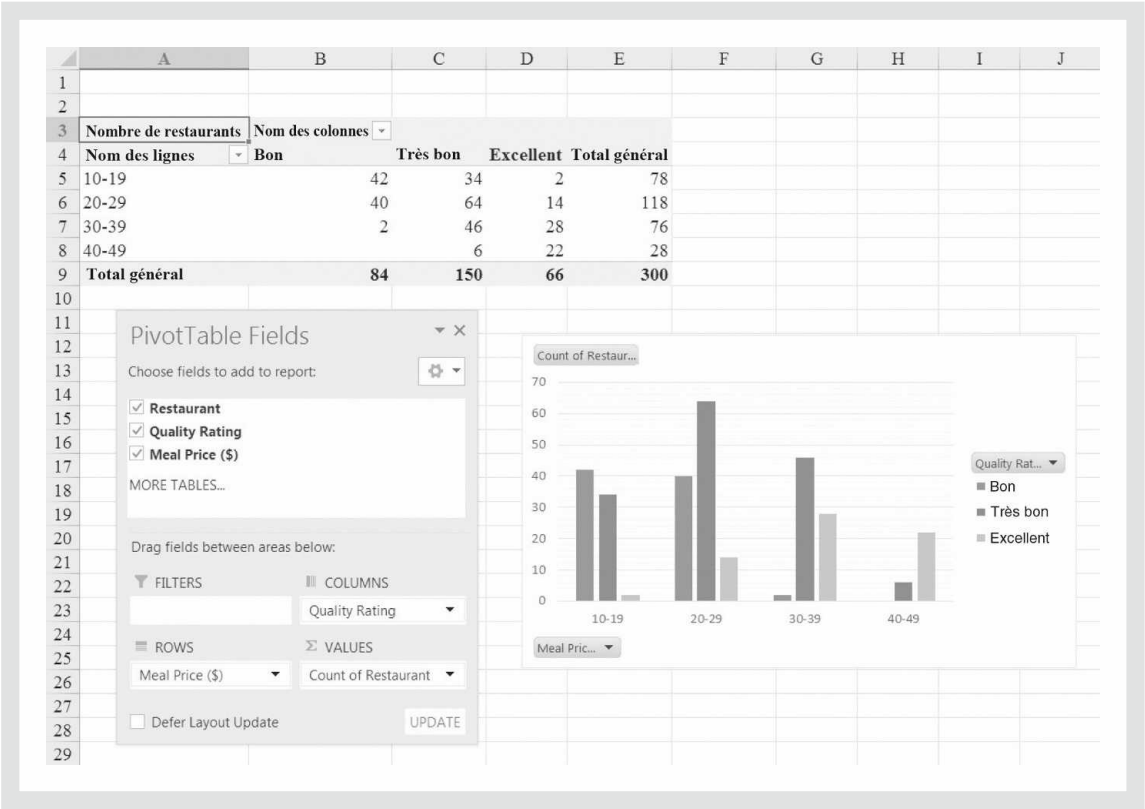


Figure 2.28 ***Diagramme en barres côte-à-côte pour les données sur les restaurants construit en utilisant l'outil Recommended Chars d'Excel***

Étape 5. Cliquer sur **Design** dans la barre des tâches (situé en-dessous du titre PivotCharts Tools)

Étape 6. Dans le groupe **Data**, choisir **Switch Row/Column** ; un diagramme en barres avec le prix des repas sur l'axe horizontal apparaît

La feuille de calcul de la figure 2.28 contient le diagramme en barres côte-à-côte pour les données des restaurants, créé en suivant ces instructions.

> Le diagramme en barres de la figure 2.28 est référencé par Excel sous le terme « Clustered Column chart ».

Options d'édition Vous pouvez aisément modifier le diagramme en barres côte-à-côte pour faire apparaître un titre de graphique différent et nommer les axes. Supposez que vous vouliez nommer le graphique « Diagramme en barres côte-à-côte », l'axe horizontal « Prix des repas (dollars) » et l'axe vertical « Fréquence ».

Étape 1. Cliquer sur le bouton **Chart Elements** + (situé à côté du coin supérieur droit du graphique)

Étape 2. Lorsque la liste des éléments du graphique apparaît :

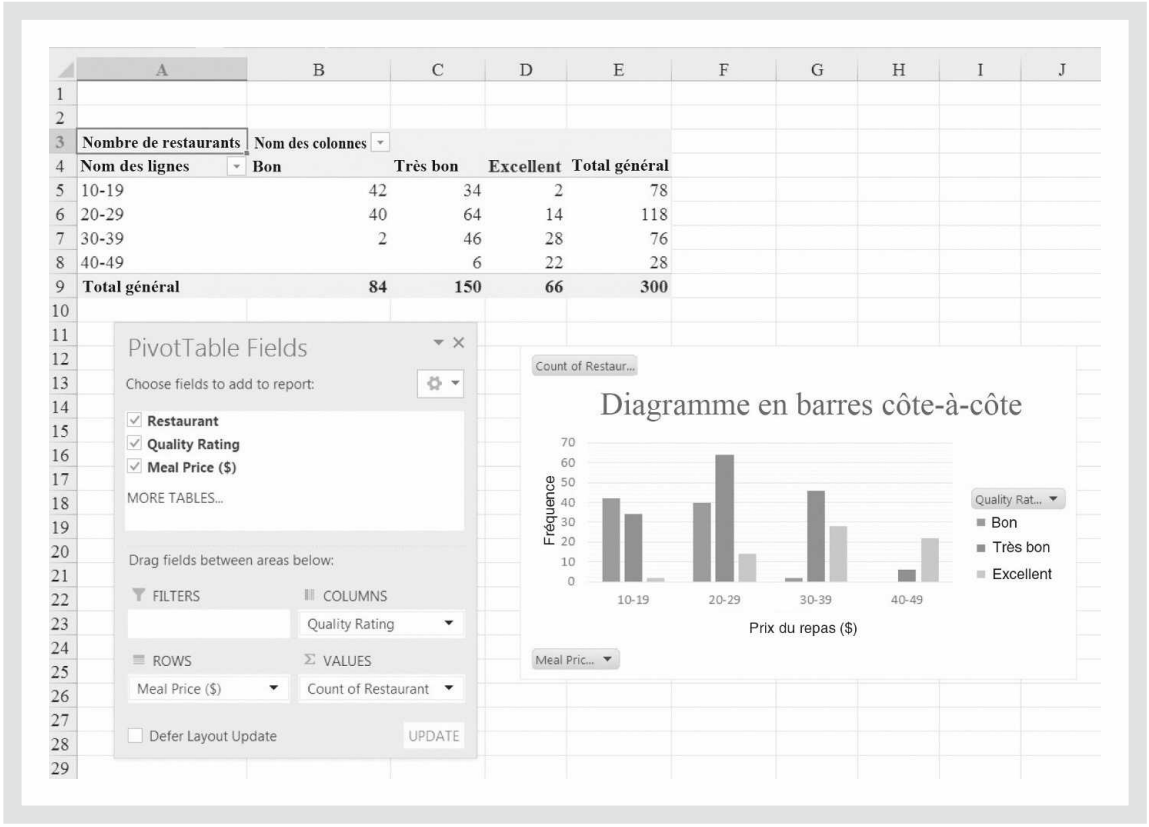


Figure 2.29 ***Diagramme en barres côte-à-côte modifié pour les données sur les restaurants construit en utilisant l'outil Recommended Chars d'Excel***

Cliquer sur **Chart Title** (crée un espace pour inscrire le titre du graphique)
Cliquer sur **Axis Title** (crée un espace pour inscrire un titre sur les axes)

Étape 3. Cliquer sur **Chart Title** et remplacer-le par **Diagramme en barres côte-à-côte**

Étape 4. Cliquer sur **Horizontal (Category) Axis Title** et remplacer-le par **Prix des repas (dollars)**

Étape 5. Cliquer sur **Vertical (Value) Axis Title** et remplacer-le par **Fréquence**

Le diagramme en barres modifié est présenté à la figure 2.29.

Vous pouvez facilement changer le diagramme en barres côte-à-côte pour obtenir un diagramme en barres empilées en suivant les étapes suivantes.

Étape 1. Cliquer sur **Design** dans la barre des tâches

Étape 2. Dans le groupe **Type**, cliquer sur **Change Chart Type**

Étape 3. Lorsque la boîte de dialogue apparaît :
Sélectionner l'option **Stacked Columns**
Cliquer sur **OK**

Une fois que vous avez créé un diagramme en barres côte-à-côte ou empilées, vous pouvez facilement passer de l'un à l'autre en répétant les deux dernières étapes.

ANNEXE 2.3 UTILISER STATTOOLS POUR CONSTRUIRE DES PRÉSENTATIONS GRAPHIQUES ET SOUS FORME DE TABLEAUX

Dans cette annexe, nous montrons comment utiliser StatTools pour construire un histogramme et un nuage de points.

A2.3.1 Histogramme

Nous utilisons pour illustrer la démarche les données sur la durée des audits du tableau 2.4 (fichier en ligne Audit). Commencer par utiliser le « Data Set Manager » pour créer un ensemble de données StatTools à partir de ces données en utilisant la procédure décrite dans l'annexe du chapitre 1. Les étapes suivantes permettent de créer un histogramme.

Étape 1. Cliquer sur le bouton **StatTools** dans la barre des tâches
Étape 2. Dans le groupe **Analyses**, cliquer sur **Summary Graphs**
Étape 3. Choisir l'option **Histogram**
Étape 4. Lorsque la boîte de dialogue apparaît :
Dans la section **Variables**, sélectionner **Durée des audits**
Dans la section **Options**,
Entrer 5 dans la boîte **Number of Bins**
Entrer 9.5 dans la boîte **Histogram Minimum**
Entrer 34.5 dans la boîte **Histogram Maximum**
Choisir **Categorical** dans la boîte **X-Axis**
Choisir **Frequency** dans la boîte **Y-Axis**
Cliquer sur **OK**

Un histogramme pour les données sur les audits similaire à celui présenté à la figure 2.5 apparaîtra. La seule différence est que l'histogramme créé en utilisant StatTools indique les centres de classe sur l'axe horizontal.

A2.3.2 Nuage de points

Nous utilisons les données sur le magasin de hi-fi contenues dans le tableau 2.14 pour illustrer la construction d'un nuage de points. Commencer par utiliser le « Data Set Manager » pour créer un ensemble de données StatTools à partir de ces données en utilisant la procédure décrite dans l'annexe du chapitre 1. Les étapes suivantes permettent de créer un nuage de points.

Étape 1. Cliquer sur le bouton **StatTools** dans la barre des tâches

Étape 2. Dans le groupe **Analyses**, cliquer sur **Summary Graphs**
Étape 3. Choisir l'option **Scatterplot**
Étape 4. Lorsque la boîte de dialogue apparaît :
Dans la section **Variables**,
Dans la colonne intitulée **X**, sélectionner Nombre de spots publicitaires
Dans la colonne intitulée **Y**, sélectionner Volume des ventes
Cliquer sur **OK**

Un nuage de points similaire à celui présenté à la figure 2.26 apparaîtra.